The Journal of

Economic Perspectives

A journal of the
American Economic Association

Fall 2020

The Journal of
Economic Perspectives

A journal of the American Economic Association

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

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

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, USA.
Printed by LSC Communications, Owensville, Missouri, 65066, USA.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Fall 2020, Vol. 34, No. 4. 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 $24.00, $34.00, or $44.00 depending on income; for an additional $15.00, you can receive this journal in print. The journal is freely available online. 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 34 • Number 4 • Fall 2020

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 3–26

The Rise of Income and Wealth Inequality in America: Evidence from Distributional Macroeconomic Accounts

Emmanuel Saez and Gabriel Zucman

For the measurement of income and wealth inequality, there is no equivalent to Gross Domestic Product statistics—that is, no government-run standardized, documented, continually updated, and broadly recognized methodology similar to the national accounts which are the basis for GDP. Starting in the mid-2010s, we have worked along with our colleagues from the World Inequality Lab to address this shortcoming by developing "distributional national accounts"—statistics that provide consistent estimates of inequality capturing 100 percent of the amount of national income and household wealth recorded in the official national accounts.

This effort is motivated by the large and growing gap between the income recorded in the datasets traditionally used to study inequality—household surveys, income tax returns—and the amount of national income recorded in the national accounts. The fraction of national income that is reported in individual income tax data has declined from 70 percent in the late 1970s to about 60 percent in 2018. The gap is larger in survey data, such as the Current Population Survey, which do not capture top incomes well. This gap makes it hard to address questions such as: What fraction of national income is earned by the bottom 50 percent, the middle 40 percent, and the top 10 percent of the distribution? Who has benefited from economic growth since the 1980s? How does the growth experience of the

■ *Emmanuel Saez and Gabriel Zucman are Professors of Economics, both at the University of California, Berkeley, California. Their email addresses are saez@econ.berkeley.edu and zucman@berkeley.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.4.3.

different groups of the population in the United States compare to that seen in other countries?

Distributing the totality of income and wealth allows us to compute income growth rates for the different social groups consistent with the official macroeconomic growth rates, thus bridging the gap between macroeconomic analysis and the study of inequality. This procedure reduces arbitrariness compared to approaches that focus on narrower notions of economic resources. In addition, because the macroeconomic aggregates are defined and estimated following harmonized, internationally agreed upon concepts and methods, distributional national accounts should maximize the comparability of inequality over time and across countries.

Piketty, Saez, and Zucman (2018) present a prototype of distributional national accounts for the United States. These series are supplemented by a set of publicly available micro-files representative of the US population. In these microfiles, each variable corresponds (and adds up) to a national account aggregate, such as compensation of employees, corporate profits, or income taxes paid; and each observation is a synthetic individual created by combining tax, survey, and other publicly available data sources. These microfiles allow anyone to reproduce all our findings on US inequality—including those described in this article—and to compute other statistics of interest. In the same way as the national accounts are constantly updated, revised, and refined, we regularly update our series and micro-files whenever new data become available and when improved estimation techniques are designed. These revisions are documented in methodological notes that explain the changes made and their effect on previously reported statistics. Following regularly updated guidelines (Alvaredo et al. 2020), similar methods are applied to construct prototype distributional national accounts in a growing number of countries, including France, India, China, and Brazil. The series are made available on the World Inequality Database at http://WID.world, along with all computer code and technical appendices. Because the code and raw data are generally publicly available, alternative methodologies can be tested.

In time, we hope that our prototype distributional national accounts will be taken over by governments and published as part of the official toolkit of government statistics. Inequality statistics are too important to be left to academics, and producing them in a timely fashion requires resources that only government and international agencies possess. A similar evolution happened for the national accounts themselves, which were developed in the first half of the twentieth century by scholars in the United States (such as Simon Kuznets), the United Kingdom (such as James Meade and Richard Stone), France (such as Louis Dugé de Bernonville), and other countries, before being taken over by government agencies.

It may take decades before we get there. Economic statistics, like aggregate output or concentration of income, are not physical facts like mass or temperature. Instead, they are creations that reflect social, historical, and political contexts. How the data sources are assembled, what conceptual framework is used to combine them, what indicators are given prominence: all of these choices reflect objectives that must be made explicit and broadly discussed. Before robust distributional

national accounts are published by government agencies, there are still many methodological choices to be debated and agreed on by the academic and statistical community. As part of that process, our prototype can be used to characterize the rise of inequality in the United States, to confront our methods and findings with those of other studies, and to pinpoint the areas where more research is needed.

The Rise of Wealth Inequality

A first step toward the creation of US distributional national accounts was taken in Saez and Zucman (2016), who produced estimates of US wealth inequality allocating 100 percent of the household wealth recorded in the Financial Accounts, the official US macroeconomic balance sheet. Household wealth includes all the nonfinancial assets (such as real estate) and financial assets (such as equities, bonds, and pension wealth, whether held in individual retirement accounts or through pension funds) of US households, net of debts. In 2019, the Federal Reserve released its own official Distributional Financial Accounts painting a similar picture of a large rise in wealth concentration.

Measuring Wealth When There Is No Administrative Data on Wealth

Because there is no administrative data on wealth in the United States, Saez and Zucman (2016) use an indirect method, known as the income capitalization technique, to estimate wealth inequality. The idea is to link the Financial Accounts aggregates to the income flows that these assets generate: thus, interest-bearing assets are linked to interest payments, corporate equities are linked to dividends and capital gains, business assets are linked to business profits, and so on. Concretely, if the ratio between the stock of interest-bearing assets in the Financial Accounts and the total flow of interest income reported in tax returns is 50, then someone with $1,000 in interest is assigned $50,000 in bonds, saving accounts, and other interest-generating assets. Wealth, in other words, is estimated by capitalizing income; in the preceding example, interest is capitalized using a capitalization factor of 50, or equivalently, an interest rate of 2 percent. Because not all assets generate taxable income (pensions, most importantly, do not), tax data need to be supplemented with other data sources to capture all forms of wealth.

The basic capitalization method is simple and transparent, and it delivers results consistent with other evidence about US wealth. In 2016, according to the basic capitalization method, US billionaires owned $3.1 trillion in wealth, a number close to the $3.0 trillion implied by the Forbes annual list of the 400 wealthiest Americans. Tax units with less than $1 billion and more than $50 million in net wealth owned $9.2 trillion, a number not dissimilar to the $10.2 trillion found in the Survey of Consumer Finances.

In its simplest form, the capitalization method relies on the assumption that within an asset class, the link between income reported in tax returns and wealth is the same across individuals; in other words, that people have the same realized rate

of return to wealth. Of course, not everybody actually has the same realized rate of return. The rates of return may even be positively correlated with wealth. In Saez and Zucman (2016), we showed that the assumption of constant realized returns within an asset class appeared reasonable, based on data from estate tax returns matched to the income tax return of the decedent the year before death, the Survey of Consumer Finances, and tax returns from foundations. In particular, we showed that the capitalization technique works well for US foundations despite the fact that the wealthiest foundations—with sophisticated investments in private equity and hedge funds—have higher total rates of returns than less wealthy foundations. The reason for this apparent paradox is that the high total returns of top foundations stem from high unrealized capital gains, not from high realized income (interest, dividends, realized capital gains) relative to wealth. What matters for the capitalization method is that within an asset class, the flow of realized income be proportional to wealth, which generally seems to be the case.

However, we also found in matched estate-income tax data, an interest rate premium that seemed to appear among the rich starting around the time of the Great Recession of 2008–2009 and noted (Saez and Zucman 2016, p. 550) that this pattern should be watched. Subsequent research suggests that the interest rate premium of the rich has become a fixture of the post-Great Recession era. In the Survey of Consumer Finances, the top 1 percent richest households have a higher-than-average interest rate in the 2010, 2013, and 2016 waves of the survey by a factor of 1.3 (Bricker, Volz, and Hansen 2018; Saez and Zucman 2019a). In matched estates-income tax data, estates above $10 million have continued to exhibit a slightly higher interest rate than average Americans after 2012, the last year in Saez and Zucman (2016) (Smith, Zidar, and Zwick 2020). Thus, assuming that all Americans have the same interest rate exaggerates the interest-bearing assets of the wealthy in the post Great Recession period.

For equity wealth, the capitalization method infers assets based on dividends and realized capital gains, and thus it cannot capture the wealth of someone who receives no dividend and barely realizes any capital gains. A striking example is given by Warren Buffett, the main shareholder of Berkshire Hathaway, a company that does not pay dividends. In 2016, Buffett disclosed he had an adjusted gross income of $11.5 million in 2015, a negligible realized return relative to the value of his stake in Berkshire Hathaway, which amounted to about $60 billion (as reported in Cohen 2016). Six of the top 10 wealthiest Americans—Jeff Bezos (Amazon), Mark Zuckerberg (Facebook), Warren Buffett (Berkshire Hathaway), Sergey Brin (Alphabet), Larry Page (Alphabet), and Elon Musk (Tesla), collectively worth more than $600 billion in September 2020, which is 0.6 percent of all US wealth—are the main shareholders of corporations that do not pay dividends. Indeed, by triangulating the available sources on the reported incomes of the ultra-wealthy, Saez and Zucman (2019a) estimate that the top 400 wealthiest Americans as a whole earn less taxable equity income (dividends and realized capital gains) relative to their equity wealth than the rest of the population by a factor of about 2. Assuming that all Americans have the same realized return on

equities thus underestimates the equity assets of billionaires—a problem that has become more acute in the 2010s with the growth of giant tech companies that typically do not distribute dividends.

Capturing these trends calls for implementing a more sophisticated version of the capitalization method. The September 2020 update of the Saez and Zucman (2016) estimates of wealth inequality, published on the World Inequality Database and also presented in this paper, incorporate the interest rate premium of the rich seen in matched estates-income tax data. They also upgrade the equity wealth of billionaires so that the total net worth of billionaire keeps matching Forbes. These changes do not significantly affect the level of top wealth shares nor their trend but bring asset composition in line with the existing evidence. For example, in these updated series, interest-bearing assets account for 23 percent of the wealth of the top 1 percent in 2018, consistent with the asset composition seen in the official Federal Reserve data on wealth inequality.[1]

Distributional Financial Accounts: A Landmark

In 2019, the Federal Reserve released its own Distributional Financial Accounts. It was the first time that the Federal Reserve published statistics on wealth covering the entire population—from the bottom 50 percent up to the top 1 percent—consistent with its own official macroeconomic balance sheets.[2] Like in Saez and Zucman (2016), the Distributional Financial Accounts start from the Financial Accounts aggregate and allocate these totals across the population.

Methodologically, the two approaches have some differences. The Federal Reserve relies on the Survey of Consumer Finances supplemented with the Forbes 400 to allocate the Financial Account aggregates; it does not use income tax data. The Survey of Consumer Finances is a high-quality wealth survey that oversamples the rich. However, the survey is only conducted triennially, starting in 1989. Thus, the Distributional Financial Accounts start in 1989 and the data is interpolated between each wave of the Survey of Consumer Finances. Like all household surveys, the Survey of Consumer Finances relies on self-reported information and suffers from small sample sizes at the top. In the latest wave of the survey, about 6,200 families were sampled.

[1] Smith, Zidar, and Zwick (2020) also modify the benchmark Saez and Zucman (2016) capitalization method. However, they assume the wealthiest Americans earn an interest rate higher than what is seen in the datasets where both income and wealth can be observed. This leads them to underestimate the interest-bearing assets of the wealthy. Smith, Zidar, and Zwick (2020) also infer equity wealth based on dividend income, despite the fact that the wealthiest Americans often own equities that do not pay dividends. As a result, they capture only 57 percent of the billionaire wealth estimated by Forbes. Once the correct interest rate is used and equity wealth is fixed to match the estimates of billionaire wealth from Forbes, the Smith, Zidar, and Zwick (2020) estimates are very close to the Saez and Zucman (2016) updated series; for discussion, see Saez and Zucman (2020).

[2] When we produced our wealth inequality estimates, we had a fruitful exchange with the researchers at the Federal Reserve who produce and analyze the Survey of Consumer Finances. These exchanges, sometimes vigorous (Bricker et al. 2016; Bricker, Henriques, and Hansen 2018), helped nurture the creation of the Federal Reserve Distributional Financial Accounts, a key and widely accessible tool.

The Federal Reserve includes consumer durables and unfunded pensions in its definition of wealth, in contrast to Saez and Zucman (2016). Although including durables and unfunded pensions can be appropriate for some purposes, it raises some issues. Durables are not assets in the UN System of National Accounts (Semega et al. 2019); other countries do not include these items in their estimates of aggregate household wealth (Piketty and Zucman 2014). Unfunded pensions—99 percent of which involves promises to government employees (in 2018)—are not backed by actual wealth. Including unfunded pensions in wealth would logically call for also including promises of future Social Security benefits and promises of other future government benefits (such as Medicare, future spending on education, and other promises net of future taxes), which neither the Federal Reserve, nor Saez and Zucman (2016), nor other countries do. For international comparability and conceptual consistency, durables and unfunded pensions are best left out of wealth.

One important but subtle issue in thinking about inequality is whether to measure the distribution of economic resources across households, as the Federal Reserve does, or across adult individuals or tax units, as in Saez and Zucman (2016). There are more tax units (180 million in 2016) than households (126 million), because roommates form separate tax units but one household, as do parents living with an adult child, and unmarried partners. We believe that data users should be allowed to choose the unit of observation that fits the question they are asking.

For instance, if one is interested in tax reforms, like the introduction of a wealth tax, then the tax unit is the proper unit of observation. In the micro-files of Piketty, Saez, and Zucman (2018), one can look at the distribution of wealth across tax units or across adult individuals with the assumption that wealth is equally split among married spouses. This "equal-split adult" approach assumes that there is a full sharing of resources between married spouses—albeit not between unmarried partners, in contrast to the household-based approach. One merit of using equal-split adults is that it improves the comparability of inequality statistics over time and across countries, because the definition of adult (in our case an individual aged 20 or more) is fixed, while definitions of households and tax units can vary. However, equal-split adult statistics understate inequality because not all wealth is equally shared among married spouses. In France, Frémeaux and Leturcq (2020) find that a growing fraction of wealth is individualized, as opposed to jointly owned between spouses. An important area for future research involves collecting more data on the division of wealth between spouses. It would also be helpful if the Federal Reserve allowed users to look at the distribution of wealth across individuals and tax units.

In the meantime, we can convert the Federal Reserve Distributional Financial Accounts from households to tax units ourselves and compare the resulting distributions to Saez and Zucman (2016).[3] Once the same unit of observation and the same

[3]For all intents and purposes high-end families are the same as high-end tax units. In Saez and Zucman (2016), "the top 1 percent" includes 1.8 million tax units in 2016, while in the Federal Reserve data, "the top 1 percent" includes 1.26 million households and around 1.26 million tax units—that is, it captures only the top 0.7 percent wealthiest tax units. Standard Pareto-interpolation techniques imply that the

definition of wealth are used, the Federal Reserve Distributional National Accounts are very close to the Saez and Zucman (2016) estimates. As shown in Figure 1, in both cases, the top 10 percent wealthiest tax units owned 77–78 percent of wealth in 2018, an increase of 10 points since 1989. In both cases, the top 1 percent wealthiest tax units owned 38 percent of wealth in 2018, also an increase of 10 points since 1989.

Overall, whether one looks at the absolute level of wealth at the top, the shares of wealth owned by the top groups, the portfolio allocation of the wealthy—and how all of this has evolved since 1989—the Distributional Financial Accounts and the Saez and Zucman (2016) estimates paint the same picture. By construction, total wealth is the same in these two datasets, equal to the Financial Accounts aggregate. In 2018, the 1 percent richest tax units had about 38 times the average wealth of $482,000 that year—that is, about $18 million on average. In terms of portfolio composition, interest-bearing assets account for close to a quarter of the net wealth of the top 1 percent in both datasets and pension assets for 10 percent.

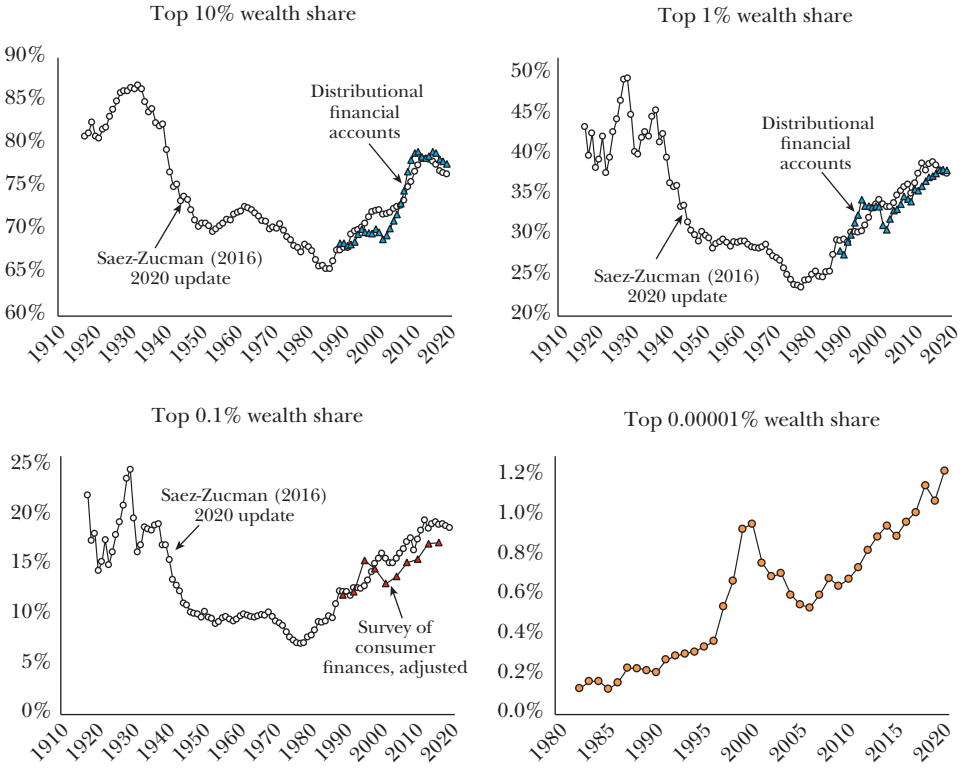
In the Distributional Financial Accounts, the Federal Reserve chooses not to report wealth statistics for the top 0.1 percent or smaller groups. But we can apply the Federal Reserve methodology and compute the top 0.1 percent wealth share in that way. As shown in Panel C of Figure 1, the Federal Reserve data again appear consistent with the Saez and Zucman (2016) estimates, although the increase in the top 0.1 percent wealth share is slightly more pronounced in capitalized income statistics. Given the limitations of the capitalization method, the Saez and Zucman (2016) series might overestimate the rise of the top 0.1 percent wealth share. But it is at least equally likely that the Survey of Consumer Finances underestimates the rise of this top share because the Survey of Consumer Finances does not capture the full extent of the rise of income inequality at the top end of the scale.[4]

The Forbes 400 ranking, which roughly corresponds to the wealthiest 0.00025 percent of households, can be used to focus on much narrower slices of the wealth distribution. These data confirm the surge of wealth concentration seen in tax data: the top 0.00025 percent wealth share according to Forbes has increased even faster than the top 0.1 percent wealth share according to the tax data. To be sure, Forbes is far from an ideal data source. It may miss people who own wealth in

share of the top 0.7 percent within the top 1 percent is $\left(\frac{0.7}{1}\right)^{\frac{\alpha-1}{\alpha}}$, where a is the Pareto coefficient, equal to about 1.3 in the Distributional Financial Accounts. Therefore, one needs to multiply the share of wealth owned by the richest 1 percent households by 1.08 to capture the share of wealth owned by the richest 1 percent tax units. Excluding consumer durables and unfunded pensions, the top 1 percent wealthiest households have 35.4 percent of total wealth in the Federal Reserve Distributional Financial Accounts in 2018, hence the top 1 percent wealthiest tax units have 38 percent of total wealth, a number identical to the Saez and Zucman (2016) estimate.

[4]Respondents to the Survey of Consumer Finances are asked about their income as reported on their tax return. But as pointed in Saez and Zucman (2016), the top 0.1 percent capital income share rose less in the SCF than in the real world tax data from 1989 to 2016. Bricker et al. (2016, p. 290) argue that this gap may owe to income misclassification: SCF respondents may, for example, call wages what in fact is business income. But the share of total income earned by the top 0.1 percent also rose less in the SCF than in the entire population, suggesting that the SCF does not capture the full extent of the rise in the top 0.1 percent wealth share.

Figure 1
Top Wealth Shares in the United States: Comparing Estimates



Source: Federal Reserve, Saez and Zucman (2016), September 2020 update, and *Forbes*.
Note: All the series use the same definition of household wealth (the market value of all non-financial and financial assets net of all debts, excluding consumer durables and unfunded pensions), have the same total wealth (the official Financial Accounts total, e.g., $76.5 trillion in mid-2016), the same totals asset class by asset class, and use the same unit of observation (tax units). To move from households to tax units in the SCF and the Distributional Financial Accounts, we assume that each tax unit within the top 1 percent corresponds to one household, and make no correction for the next 9 percent. To make the SCF comparable to the other two sources, we add the Forbes 400 to the public-use SCF files and adjust reported wealth to match the Financial Accounts totals asset class by asset class.

diversified portfolios of stocks and bonds (for which no public information exists) and overstate the value of private businesses. To alleviate some of these concerns, we can focus on the very top of the ranking, the top 0.00001 percent wealthiest Americans, a group that includes 17 tax units today and 10 in 1982, the first year that the Forbes 400 was published. It is not unreasonable to assume that in a given year the 10 or 20 wealthiest people in the country are correctly identified by Forbes and their holdings are broadly accurately estimated. This group is a mix of major shareholders of big, publicly listed companies (in 2020, Amazon, Facebook, Google, Walmart, Microsoft, Berkshire Hathaway; in 1982, Getty Oil, Standard Oil, Hewlett

Packard, and others) for which valuations are observable and giant private businesses (Koch Industries and Bloomberg LP today; Mars and Hunt Oil Company in 1982) that attract public scrutiny. As shown by Panel D of Figure 1, the share of wealth owned by this elite group has risen from 0.13 percent of total US wealth in 1982 to 1.2 percent in 2020, an almost tenfold increase.

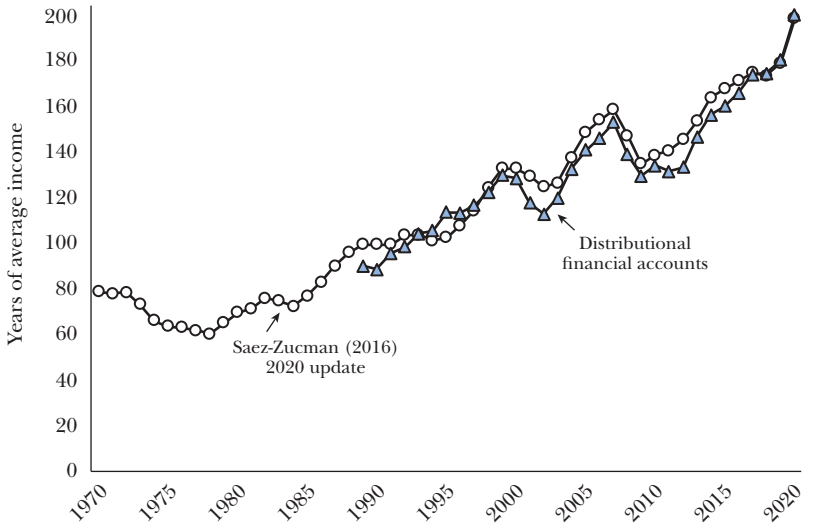
By any metric, the period from 1980 to 2020 has been an era of extraordinary wealth accumulation among the rich in the United States. Not only has wealth become more concentrated, wealth itself has been growing faster than income and output. In 1980, the ratio of aggregate household wealth to national income was 300 percent. In 2020, this ratio approaches 570 percent, the highest level ever recorded in the history of the United States. In other words, during the 1980–2020 period, wealth as a whole has been growing almost twice as fast as income. The result is that relative to what is produced and earned in a given year, the wealth of the rich has skyrocketed. In 1980, on average, members of the top 1 percent owned in wealth the equivalent of 60 years of average US income. In 2020, whether one looks at the Saez and Zucman (2016) or Distributional Financial Accounts estimates, they own 200 years of average US income in wealth, as shown in Figure 2.

Although it is notable that the main sources used to estimate US wealth inequality deliver consistent results, it would be a mistake to exaggerate our ability to measure top-end wealth. Changes in tax avoidance, the growth of wealth held in foundations, and the globalization of wealth management pose formidable challenges (for discussion, see Zucman 2015). It is a failure of public statistics that the only information on billionaire wealth comes from magazines. We could and should do better to measure wealth inequality than rely on a survey of 6,200 families or indirectly infer asset ownership based on income flows.

One merit of a well-administered wealth tax is that it would provide better information on the distribution of wealth, one of the most hotly debated issues in democratic societies. Even without a wealth tax, governments could collect information on assets and debts from third parties (banks, pension funds, brokers, and others), as they already do for income. These data could be used to improve tax enforcement—as currently done in Denmark—and allow for the construction of more accurate Distributional Financial Accounts.

Like all important economic statistics, the Financial Accounts themselves have limitations and remain, decades after their creation, a work in progress. One challenge involves the valuation of private business assets, which tends to be conservative in these accounts. Another relates to offshore wealth: foreign bank accounts, portfolios of equities and bonds held through foreign financial institutions, and holdings of foreign mutual funds (including hedge funds) that are not intermediated through a US broker, are not captured in the Financial Accounts (Zucman 2013). The forms of wealth that are broadly shared tend to be accurately measured, while the more complex investments, involving legal and financial intermediaries in foreign countries with a great deal of secrecy, are less well captured. The estimates of wealth concentration we have today, which by construction are anchored to the Financial Accounts totals, should be seen as lower bounds.

Figure 2
Average Wealth of 1 Percent Wealthiest Adults
(Divided by Average US Income Per Adult)



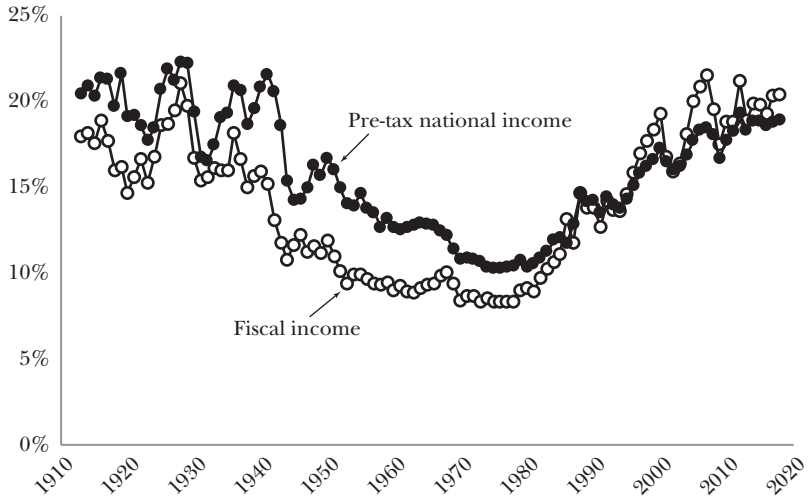
Source: Saez and Zucman (2016), September 2020 update available on WID.world, and Federal Reserve Distributional Financial Accounts.
Note: This figure shows the average wealth of the top 1 percent wealthiest adults (with wealth equally split among married spouses), expressed as a ratio to average US national income per adult. For the Distributional Financial Accounts, we assume that the average wealth of the top 1 percent households is the same as the average wealth of the top 1 percent equal-split adults.

The Rise of Income Inequality: Beyond Tax Data

Bridging the Gap between the Study of Inequality and Macroeconomics

There has been a surge of research in recent years to which we have contributed our share, measuring income inequality using administrative tax data rather than self-reported household survey data. This work has made us aware of the large and growing gap between national income and taxable income. On the labor side, untaxed labor income includes tax-exempt employment benefits (contributions made by employers to pension plans and to private health insurance), employer payroll taxes, the labor income of non-filers, and unreported labor income due to tax evasion. The fraction of labor income which is taxable has declined from 80 to 85 percent in the post-World War II decades to just under 70 percent in 2018, due to the rise of employment fringe benefits—in particular the rise of employer contributions for health insurance, particularly expensive in the United States. Most studies of wage inequality ignore fringe benefits even though they are a large and growing fraction of labor costs. As for capital, only one-third of total capital income

Figure 3
Share of Income Earned by the Top 1 Percent



Note: This figure compares the share of fiscal income earned by the top 1 percent tax units (from Piketty and Saez 2003, updated series including capital gains in income to compute shares but not to define ranks, to smooth the lumpiness of realized capital gains) to the share of pre-tax national income earned by the top 1 percent equal-split adults (from Piketty, Saez, and Zucman 2018, updated September 2020, available on WID.world).

is reported on tax returns. Untaxed capital income includes undistributed corporate profits, the imputed rents of homeowners, capital income paid to pension accounts, and dividends and interest retained in trusts, estates, and fiduciaries.

Piketty, Saez, and Zucman (2018) estimate the distribution of 100 percent of national income by combining national accounts, tax, and survey data. As Figure 3 shows, in both fiscal income and national income statistics, the share of income earned by the top 1 percent was high before the 1930s and fell from the 1930s to the 1970s before rising again from the late 1970s on. This U-shaped evolution of income concentration is a bit less spectacular when one looks at national income rather than fiscal income, mainly because only the fraction of corporate profits paid out as dividends are included in fiscal income statistics, while all corporate profits are included in national income. Accounting for the totality of corporate profits generally increases the top 1 percent income share, but the effect is stronger in the post-World War II years, a time before the rise of pension plans somewhat broadened equity ownership.

One virtue of distributional national accounts is that they are not affected by legal changes in business organization. In the United States, a growing number of businesses have been organized as "pass-through" entities since the late 1980s. The income of pass-through entities—partnerships, S-corporations, sole proprietorships—is not subject to the corporate income tax; instead, all the income of these

businesses is passed to their individual owners and subject to the individual income tax only. When more businesses operate as pass-throughs, more income mechanically shows up on individual income tax returns, especially at the top-end of the income distribution. In our distributional national accounts, all corporations are de facto treated as pass-through entities, no matter their legal status. In the same way as partnership income is allocated to partners, all corporate income is allocated to shareholders. In the same way as partners pay the individual income tax on their share of partnerships' income, shareholders pay the corporate tax on their share of corporations' income. This seems a logical way to allocate the corporate tax.

Because there is no administrative data in the United States on the ownership of non-pass-through corporations, we must make assumptions to allocate the portion of corporate profit that is not paid out as dividends. In our distributional national accounts, we allocate 50 percent of undistributed profits proportionally to dividends and 50 percent proportionally to realized capital gains. This method is far from ideal. In the real world, some people with little dividends and realized capital gains are major shareholders of corporations with large undistributed profits. For a deeper understanding of income inequality, the government should collect information about the ownership of corporations. This information exists in private financial institutions, such as the Depository Trust Company, the central securities depository of the United States, which de facto acts as the ultimate bookkeeper for the ownership of securities.[5]

From Macroeconomic Growth to People's Growth

An advantage of distributing the totality of national income is that it allows for apples-to-apples comparisons of inequality across countries because national income is defined and computed in the same way internationally. (National income is equal to GDP minus capital depreciation plus net income received from abroad.) Our benchmark statistics use the equal-split adult as the unit of observation. Our benchmark definition of income, pre-tax national income, includes all pension income (from Social Security and private pensions) and subtracts all corresponding pension contributions, making estimates comparable across countries with different age structures.

For example, compare France and the United States. In the United States, national income reached $17.5 trillion in 2018, close to $72,500 on average among the 242 million adults who lived in the United States. The bottom 50 percent earned 12.5 percent of national income, which means that members of the bottom 50 percent earned one-quarter of the average income in the economy or about

[5] In a paper titled "Capitalists in the 21st century" Smith et al. (2019) find that "the typical top earner derives most of her income from human capital." They obtain this finding by noting that pass-through business income is a key source of income at the top of the fiscal income distribution and estimating that 75 percent of this income is labor income rather than capital income. However, fiscal income misses two-thirds of total capital income, in particular the profits of corporations that do not pay dividends (such as Amazon, Facebook, and Google). Moreover, the estimate that only 25 percent of pass-through business income is capital income is not consistent with the large capital stock of these businesses. Saez and Zucman (2020) discuss these points and estimate that capital income is slightly more than half of income for the top 1 percent and about two-thirds for the top 0.1 percent of the national income distribution.

$18,500 on average. In France, using purchasing power parity exchange rates to convert euros into US dollars, national income per adult was $53,000—substantially less than in the United States. But the bottom 50 percent earned 22.5 percent of national income or about $23,400 on average. Even though average income is 37 percent higher in the United States than in France, the market delivers higher incomes to the bottom 50 percent in France than it does in the United States. The French welfare state is not responsible for this feat, as we are talking here about pre-tax national income (before government taxes and transfers other than Social Security). Moreover, the income comparison does not include the better health outcomes and more extensive leisure time in France.

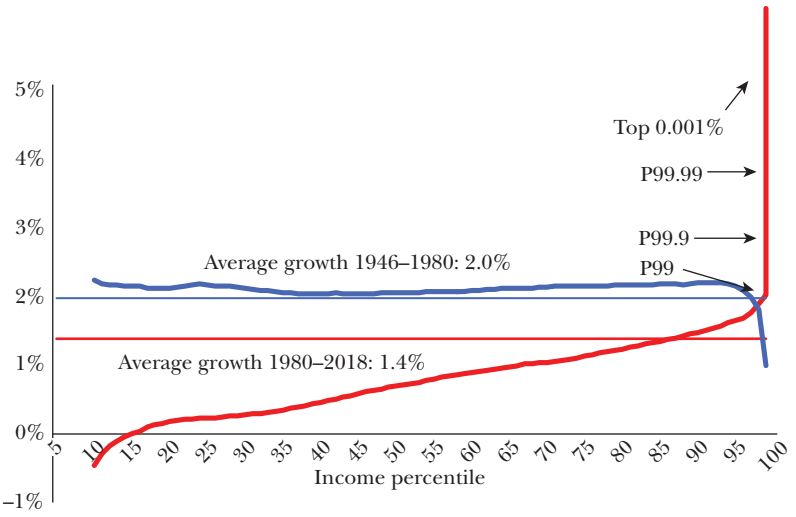
Distributing the totality of national income also allows for rigorous comparisons of income over time. Figure 4 shows the growth rate of income for each percentile of the income distribution from 1946 to 1980 and 1980 to 2018. From 1946 to 1980, average per adult national income rose 2 percent a year, one of the highest growth rates recorded over a generation in a country at the world's technological frontier. Moreover, this growth was widely shared, with only the income of the top 1 percent growing a bit less than average. One easily understands why many economists chose during this period to treat the US distribution of income as a constant.

From 1980 to 2018, average annual growth in per adult national income falls to 1.4 percent a year. For almost 90 percent of the population, growth has been below—often much below—that figure. For the bottom 50 percent as a whole, growth in pre-tax income has been only 0.2 percent on average per year. This quasi-stagnation is not due to population aging, since pre-tax income includes Social Security and other retirement benefits. Excluding the elderly (aged 65 or more), the average bottom 50 percent pre-tax income has slightly declined since 1980. During the last four decades, macroeconomic growth has not been representative of the growth experience of the vast majority of the population.

We need a different measure of economic growth to capture the lived reality of growth in an era of rising inequality. Saez and Zucman (2019b) propose "people's growth," which is the arithmetic average of the growth rate of each percentile of the income distribution. People's growth captures how income grows on average across people, as opposed to how the average income grows. From 1946 to 1980, people's growth and national income growth coincided in the United States (2.0 percent a year). From 1980 to 2018, people's growth has been only 0.65 percent a year, much less than macro growth (1.4 percent).

With a full picture of the distribution of national income over time, we can ask how income would have grown across the income distribution if growth had been equitably distributed. If macro growth had been equitably shared from 1980 to 2018, the average pre-tax income of the bottom half of the income distribution would have been 57 percent higher in 2018 than it was in actual fact. For the middle-class—from the 50th to the 90th percentile of the distribution—average incomes would have been 16 percent higher in 2018. However, for the upper middle class (from the 90th to 99th income percentile), average incomes would have been 8 percent lower, and for the rich (the top 1 percent), 36 percent lower.

Figure 4
Average Annual Income Growth Rates



Source: Saez and Zucman (2019b).
Note: This figure depicts the annual real pre-tax income growth per adult for each percentile in the 1946–1980 period (in blue) and 1980–2018 period (in red). From 1946 to 1980, growth was evenly distributed with all income groups growing at the average 2 percent annual rate (except the top 1 percent which grew slower). From 1980 to 2018, growth has been unevenly distributed with low growth for bottom income groups, mediocre growth for the middle class, and explosive growth at the top.

To be sure, this counterfactual analysis has its limitations. Perhaps with less inequality, average growth might have been lower (there would perhaps have been less innovation if million-dollar earners had not been able to earn the sums they did) or higher (there might have been more innovation if credit-constrained households had been able to earn more than they did). But the counterfactual does illustrate vividly the shift in income distribution.

Pitfalls of Personal Income Distributions

In March 2020, the US Bureau of Economic Analysis released a prototype distribution of personal income—one of the aggregate measures of income used in the US national accounts. These data provide an important step toward the creation of official distributional national accounts. But there are strong reasons to prefer national income over personal income.

First, personal income is specific to the US national accounts. It is not computed in other countries and in fact does not exist in the UN System of National Accounts. This makes it impossible to compare inequality internationally.

Second, personal income is a mixture of pre-tax and post-tax income, and for that reason not a satisfactory definition of income conceptually. Personal income does not subtract payroll taxes or individual income taxes, but it includes all individualized government transfers, such as Social Security benefits, welfare assistance, Medicare, and Medicaid. Therefore, personal income double counts some forms of income.

Third, personal income does not include corporate profits; it only includes the portion of corporate profits distributed as dividends. As a result, personal income is affected by businesses' choices of organizational form. If a person operates as a pass-through entity, all of that person's income gets counted in personal income. If the same person operates as a corporation, her income can be zero. Warren Buffet has billions in pre-tax national income; his personal income is smaller by a factor of 1,000. Unsurprisingly, the inequality of personal income is lower than the inequality of national income. If more individuals incorporate to take advantage of the low federal corporate tax rate enacted in 2018, personal income and its concentration will fall, even though nothing else will change than the tax form used by the business owner. The distribution of personal income is likely to become a poorer and poorer indicator of income inequality.

The Bureau of Economic Analysis (BEA) justifies the choice of personal income by stating that this aggregate "is closest to the measure of economic resources available to households to purchase goods" and that "[s]tarting with personal income will allow further analysis of disposable personal income (after taxes) and a better comparison to consumption" (Fixler, Gindelsky, and Johnson 2020, p. 3). The implicit view is that consumption is what matters. Consistent with this view, the BEA uses the household as the unit of observation, not the adult individual as in Piketty, Saez, and Zucman (2018).[6] Our own view is that income and consumption both deserve to be studied, but separately, because they are distinct concepts.

National income is a more meaningful concept to study income inequality than personal income because it includes all the forms of income that accrue to individuals no matter the specific ways in which this income is earned, consumed, or saved. The notion of personal income was popular among BEA statisticians in the 1950s; the first distributions of personal income were computed at that time. In the 1950s, when large corporations were controlled by multiple stakeholders, what happened in the realm of corporations could feel disconnected from what happened in the realm of households. Today, shareholders exercise much more control over their firms; the frontier between corporations and households is fuzzy. The fiction that what happens in the corporate world has nothing to do with income inequality is no longer tenable. Looking forward, it is essential for the BEA to distribute national income.

[6]The BEA also uses household equivalence scales, in which household income is divided by the square root of household size, as an adjustment for differences in household size, which makes it impossible to draw direct connections from distributions to macroeconomic growth.

How Government Taxes and Spending Affect Inequality

In the United States, federal, state, and local governments collect about 28 percent of national income in taxes and spend more than 28 percent of national income. In Denmark and France, taxes and government spending reach 50 percent of national income. Our distributional national accounts can be used to ask questions like: Do high-income people pay more or less in taxes relative to their income than the average individual? How do cash transfers compare to taxes for low-income groups? Are middle-class incomes higher after taxes and transfers than before taxes and transfers?

According to a widespread view, a government transfer is simply a tax with a minus sign, and all that matters is people's budget sets, net of all taxes and transfers. We emphasize, however, that taxes and transfers are distinct objects that must be studied as such. For example, taxes reduce cash income but most government transfers do not increase cash income. The bulk of government transfers are in-kind—such as Medicaid health insurance for the poor and Medicare for the elderly—or take the form of collective consumption, such as spending on education, police, and defense. Even when transfers are monetary, it's important to recognize that taxes are often paid cash on the nail, while transfers are generally received with a lag. For a poor, credit-constrained worker, paying $100 in payroll taxes each and every month and receiving a one-time check of $1,200 the following year (such as an Earned Income Tax Credit payment) is not equivalent to zero tax and no transfer. With an extra $100 a month, people are less likely to default on a monthly rent or interest payment; they are more likely to be able to afford an emergency expense, such as a visit to the doctor, and to afford basic daily consumption needs, such as food for their families.

The Collapse of Tax Progressivity

There is a long tradition of research on the distribution of US taxes, pioneered by Colm and Tarasov (1940), Musgrave et al. (1951), and Pechman and Okner (1974). This tradition has been refined by government agencies. Our distributional national accounts make four main departures relative to the analysis carried out by US agencies—most prominently the Congressional Budget Office—and think-tanks.

First, we include taxes at all levels of government, instead of federal taxes only. State and local taxes are sizable: about 10 percent of national income, one-third of total tax revenue. In addition, state and local governments often make substantial use of sales and excise taxes that are regressive. And so, ignoring these taxes gives a misleading picture of the progressivity of the tax system.[7]

Second, we consider taxes as a share of pre-tax national income, the broadest and most consistent definition of income. This is particularly important because a sizable fraction of the true pre-tax income of the wealthy—their share of corporate profits that is not paid out as dividends or realized as capital gains in a given year—is not

[7] The Institute on Taxation and Economic Policy (2018) is the only institution that provides comprehensive distributional state and local tax analysis state by state.

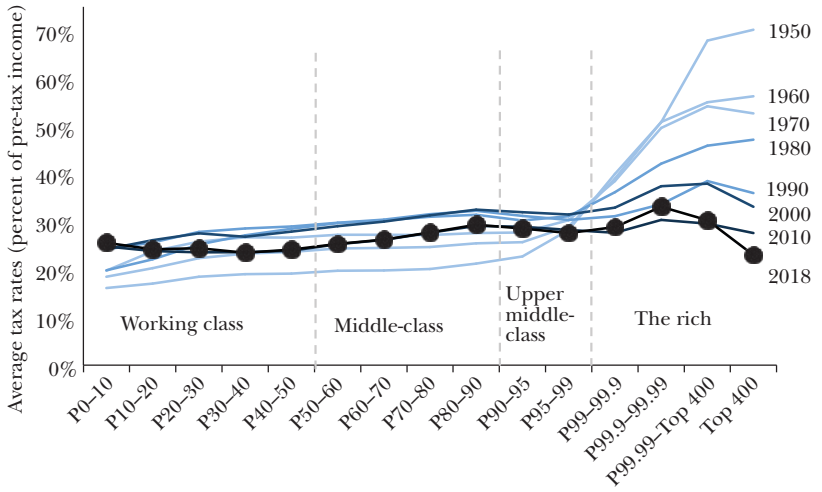
subject to individual income taxation. Because we include all taxes and all national income in our analysis, the average tax rate in our statistics is equal to the officially recorded macroeconomic tax rate, 28 percent of US national income in 2018.

Third, we do not shift taxes from one factor of production to another. In our statistics, consumption taxes are assigned to consumers, labor taxes are assigned to the corresponding workers (even when employers nominally pay them), and capital taxes are assigned to the corresponding owners of capital. In particular, the corporate tax is assigned to shareholders, just like the income tax paid on the profit of pass-through businesses is assigned to the owners of pass-through businesses. This framework allows us to allocate all taxes while keeping national income constant in a conceptually consistent manner (as discussed in Saez and Zucman 2019c). It also makes it possible to measure the economically relevant tax wedge on each factor of production, such as the gap between what it costs to employ a worker and what the worker receives. For the most part, the methodology currently followed by government agencies to study the distribution of taxes is similar to the methodology we use. It allocates all labor taxes to workers, all consumption taxes to consumers, most capital taxes to capital owners, and keeps national income constant. However, it shifts part of the corporate tax to people other than shareholders. The corporate tax is partly allocated to workers because it is assumed to depress domestic capital and reduce wages. This procedure is inconsistent with maintaining a constant level of national income and leads to biased trends in tax progressivity.[8]

Finally, our analysis treats refundable tax credits as government transfers—not as negative taxes. In the national accounts, payments made by the government to people are transfers, no matter which administration is in charge of sending these transfers. That the Earned Income Tax Credit (EITC) payments are administered by the Internal Revenue Service (rather than, say, the Social Security Administration) has no economic implication. In the macroeconomic statistics of tax revenues (for instance, the government revenue statistics published by the OECD), the refundable portion of the EITC is never subtracted from taxes. The same is true for the refundable portion of the Child Tax Credit. Proponents of the EITC felt that the program would be more acceptable politically if presented as a tax reduction rather than a transfer, and a large portion of the US public has become used to thinking about the EITC as a negative tax. The Congressional Budget Office and some think-tanks that produce distributional tax statistics choose to subtract refundable tax credits from taxes paid. But what may be perceived as good politics does not necessarily correspond to what is most conceptually consistent. Economically, the EITC is no different from other cash transfers to low-income families.

[8]For example, the Congressional Budget Office (2012) allocates 25 percent of the corporate tax to workers and 75 percent to capital owners, including owners of interest-bearing assets. If a C-corporation elects to be treated as an S-corporation (a pass-through business), then in the CBO treatment pre-tax income inequality increases (income that was previously assigned to workers is now allocated to shareholders, who are higher up in the income distribution), the labor share of national income falls, and the tax system becomes more progressive (taxes that used to be paid by workers are now paid by shareholders), despite the fact that nothing real has changed in the economy or in the tax system.

Figure 5
Average Tax Rates By Income Groups
(percent of pre-tax income)



Source: Saez and Zucman (2019b).
Note: The figure depicts the US average tax rate by income groups from 1950 to 2018. All federal, state, and local taxes are included. Taxes are expressed as a fraction of pre-tax income. P0-10 denotes the bottom 10 percent of the income distribution, P10-20 the next 10 percent, etc.

The choices we make in our distributional national accounts are of course not the only possible ones, but we stress that they are the opposite of arbitrary. Instead, they follow consistent, internationally defined economic concepts. Using concepts that are the product of international deliberation—at least as a starting point—can help control the effect of national political and ideological idiosyncracies and contribute to more coherent and comparable statistics.

When taking a comprehensive perspective on taxation, a dramatic decline in the progressivity of the US tax system appears. Figure 5 depicts the US average tax rate by income groups for various years from 1950 to 2018. All federal, state, and local taxes are included and taxes are expressed as a fraction of pre-tax income. P0-10 denotes the bottom 10 percent of the income distribution, P10-20 the next 10 percent, and so on. We split the top 10 percent into smaller groups all the way to the top 400 wealthiest Americans popularized by *Forbes*. Taking all taxes together, the US tax system used to be slightly progressive or roughly proportional for the bottom 99 percent of the income distribution but highly progressive within the top 1 percent. In 1950, for example, the upper middle class (the top 10 percent excluding the top 1 percent) paid average tax rates of around 25 percent, while the top 0.01 percent paid almost 70 percent of its income in taxes.

The tax system was highly progressive in the 1950s because corporate profits, the main source of income for the rich, were subject to a high effective corporate tax rate of 50 percent. Very high top marginal individual income tax rates (91 percent until 1963 at the federal level) made it impossible for business owners to bypass the corporate tax by using pass-throughs, such as partnerships. Moreover, high incomes were hit both by the progressive individual income tax on their realized capital income and by a progressive estate tax at the time of death. The combination of the income tax, the corporate tax, and the estate tax made the tax system extremely progressive and hard to avoid (Saez and Zucman 2019a). Low-income households paid lower taxes than today because the payroll tax was lower in the past.

In 2018, the US tax system looks like a giant flat tax that becomes regressive at the very top end. The working class and the middle class pay substantial taxes because payroll taxes are large and state and local sales and excise taxes are regressive. The very top pays low effective tax rates because of the demise of the federal corporate tax, which in 2018 collected only 1.5 percent of national income, down from 5–7 percent in the 1950s. The effective individual income tax rate falls at the top end because the very rich earn income through corporations and can avoid reporting individual income. The regressivity of the tax system at the extreme top end in 2018 is striking—a direct consequence of the 2018 cut in the corporate tax. But the figure shows a decades-long shift, with a slow erosion of the corporate tax, the estate tax, and gradually lower progressivity of the individual income tax at the top.

If the low corporate tax of 21 percent set in 2018 continues, there is a real risk that the wealthy will incorporate, earn income through their corporations, and bypass the progressive individual income tax by retaining earnings within their corporations. If held until death, the capital gains generated by such retained earnings will never be taxed.

Have Government Transfers Offset the Rise of Inequality?

Taxes are only half of the government equation. On the spending side, Social Security (retirement and disability) and unemployment insurance replace lost labor earnings due to retirement, disability, or unemployment. These programs grew fast after World War II to about 6 percent of national income in the late 1970s and have been stable afterwards. We include these transfers in our measure of pre-tax income. The remaining forms of government spending are part of post-tax income (but not pre-tax) and can be classified in three categories, from the easiest to allocate to individuals to the hardest: cash transfers, in-kind transfers, and collective consumption. Cash (or quasi-cash) individualized transfers include welfare assistance and refundable tax credits for low-income families with children, food stamps for the poor, and Supplemental Security Income for the low-income elderly and the disabled. These transfers come closest to pure redistribution as individuals can freely (except in the case of food stamps) choose how to spend them, just like earned income. Cash transfers are small, 2–3 percent of national income with no clear trend after the mid-1970s. More specifically, refundable tax credits have grown but welfare

assistance has fallen in the same proportion. Cash transfers go overwhelmingly to the bottom 50 percent.

Next, in-kind individualized transfers, such as public health insurance (Medicaid and Medicare), housing assistance, and higher education tuition subsidies, have grown from almost zero in 1960 to about 8 percent of national income in 2018. This growth is overwhelmingly driven by Medicare and Medicaid, which account for over 90 percent of all in-kind transfers in 2018. In our distributional national accounts, we allocate these transfers as a lump sum per beneficiary; that is, we divide the sums paid on Medicare by the number of Medicare beneficiaries and assign each beneficiary the average Medicare transfer. A large fraction of in-kind transfers go to the bottom 50 percent.

The last category of government spending is collective consumption expenditure. This category includes government spending on education, defense, public order (police, prisons, courts), and other public goods. Collective consumption is large (about 18 percent of national income) and has been fairly stable since 1960. Spending on defense has shrunk while other forms of collective consumption have increased; spending on prisons has increased particularly fast, due to a massive increase in incarceration rates in the 1980s and 1990s. Government spending on education has been stable at 5 percent of national income since 1970. In our distributional national accounts, we allocate all collective consumption neutrally, so that collective consumption does not affect income inequality. Obviously, if we were to allocate collective consumption on a per capita basis, that would make inequality look lower.[9]

How does incorporating taxes and transfers affect the distribution of income? In the big picture, the tax system is approximately a flat tax—taxes are proportional to income—while the transfer system is closer to a flat amount per person. This combination reduces inequality: post-tax inequality is less than pre-tax inequality.

With our distributional national accounts, we can also examine whether changes in government intervention in the economy have lifted incomes at the bottom. The short answer is "yes, but not a lot." The average pre-tax income for the bottom 50 percent, as we have seen, has almost stagnated since 1980 in real terms: it was $17,500 per adult in 1980 and is $18,500 in 2018. After deducting taxes and adding all forms of government spending, average post-tax income has increased by 25 percent since 1980. This is better than quasi-stagnation but still less than the 70 percent increase in average income per adult from 1980 to 2018. The rise of government transfers to the bottom has offset roughly one-third of the growth gap between the bottom half and the average American.

Moreover, most of the growth in bottom 50 percent post-tax incomes is driven by the surge in Medicaid and Medicare. To see this, it is useful to consider a narrower definition of post-tax income—disposable cash income. Disposable cash

[9]The main reason why Auten and Splinter (2019) find low top income shares on a post-tax basis is because they allocate half of government consumption per capita. See Saez and Zucman (2020) for a complete discussion of Auten and Splinter (2019).

income is pre-tax income minus all taxes plus all cash or quasi-cash transfers; it excludes in-kind transfers and collective consumption expenditures. This notion of disposable income is close to the one used to measure the poverty rate (Semega et al. 2019), with the important difference that we deduct all taxes and add refundable tax credits and food stamps. Economy-wide, cash disposable income per adult has increased about as much as national income from 1980 to 2018, by close to 70 percent (thanks in part to being bolstered by growing federal deficits). Figure 6 shows that for the bottom 50 percent, disposable cash income has grown very modestly over the last four decades: it was $16,000 in 1980 and $18,600 in 2016—a 16 percent increase over 36 years.

Until 2008, the bottom 50 percent paid more in taxes than it received in cash transfers: pre-tax income was higher than cash disposable income. The cash disposable income of the bottom 50 percent of adults was lifted up by the large government deficits run during the Great Recession. Since 2012, cash disposable income is almost identical to pre-tax income. Thus, the gains in post-tax income for the bottom 50 percent over this time take the form of in-kind transfers (primarily Medicaid) and collective public expenditures (education, defense, police, and prisons being the main items).

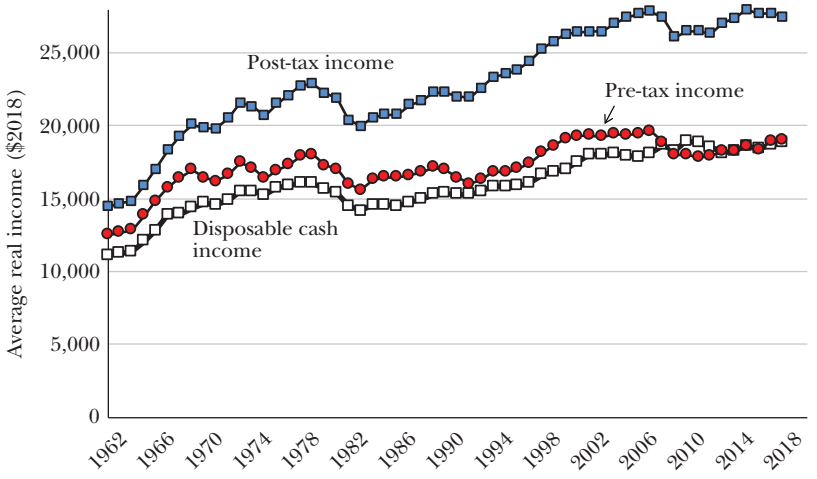
The Limits of Post-Tax Income

The modest gains in post-tax income for the working class must be analyzed with care because allocating in-kind transfers and collective consumption to individuals based on their cost for the government is highly problematic. All OECD countries have decided that everybody should have access to quality education. All OECD countries except the United States have a national program for financing health care. The cost of universal provision of education and health looks like a large transfer relative to income for low-income families. But it is conceptually incorrect to treat this full transfer as income for its recipients. After all, if low-income families received an equivalent amount in cash, most of them would not spend it all on health or education. Perhaps the best conceptual alternative would be to assign the perceived cash value of individualized in-kind transfers to recipients, while treating the rest as a collective public good.

These conceptual problems are particularly thorny for health transfers in the United States. Medicaid transfers are large because the cost of health care is extraordinarily high in the United States. The money is not flowing into the bank accounts of beneficiaries; instead, it is flowing to the bank accounts of health care providers, many of which are near the top of the income distribution. What sense does it make to rejoice in the rise of working-class post-tax incomes if this rise reflects the rise in the rents earned by the medical and pharmaceutical sectors?

A similar issue arises with government mandates, like the rule in the Affordable Care Act in 2010 that employers with 50 or more employees are legally required to provide health insurance to their full-time workers (or pay a penalty of $3,000 per employee). The cost of this mandatory private health insurance is large and growing; it is a heavy burden on low-paid workers. In conceptual terms, part of

Figure 6
The Evolution of Bottom 50 Percent Incomes



Source: Piketty, Saez, and Zucman (2018), updated September 2020.
Note: The figure depicts the evolution of the real incomes per adult (in 2018 dollars) for the bottom half of the income distribution for three income concepts: (1) pre-tax income before deducting taxes or adding government transfers (concept sums up to national income), (2) post-tax income that deducts all taxes and adds all transfers (cash and in-kind) and collective public expenditures minus the government deficit (also sums up to national income), (3) disposable cash income which is pre-tax income minus all taxes plus cash (or quasi-cash) transfers, i.e., (3) does not include in-kind transfers (primarily Medicaid and Medicare) and collective public expenditures that are included in (2).

this cost should be considered as a tax on workers that the government imposes to achieve wider health insurance coverage (Saez and Zucman 2019b). Like other taxes, this cost should be subtracted from income for the computation of post-tax income.

In short, there is no perfect measure of post-tax income. To measure the inequality of income after taxes and transfers, disposable cash income is perhaps the most meaningful concept. Disposable cash income captures income available for saving and consumption, excluding the collective consumption of services like education and health mandated by the government. But disposable cash income does not add up to national income. Post-tax national income captures all of national income by deducting all taxes and adding back all forms of government spending and the government deficit. But computing post-tax national income requires assigning collective consumption expenditures as well as the current government deficit to individuals. There is no obvious, universally "correct" way to do such an imputation, and there will never be.

Does this mean that we cannot know what is happening to inequality? Of course not. There are no raw facts in the social sciences. Rather, there are attempts at

describing reality through more or less elaborate statistical frameworks. The results of these attempts can only be properly understood once we know how the measurement tools work, what aspects of reality they aim to capture, what led to their creation, the objectives of their creators, the knowledge they embody, the accountability of the institutions that publish them, and the theories that underpin them.

Once we understand how distributional national accounts are constructed, a reasoned use of these statistics becomes possible—just like a reasoned use of GDP statistics becomes possible once we understand their strengths and limitations.

Pre-tax national income, which captures income earned from market activities, can be used to decompose macroeconomic growth and to compare inequality over time and across countries. Cash disposable income can be used to study the income available for saving and private consumption; by subtracting the saving component, it can be used to study consumption inequality. Post-tax national income can be used to estimate the total direct distributive effects of government intervention in the economy. All of these notions have merits and demerits and must be studied jointly. Ultimately, the best data would be published by government agencies, accountable to elected representatives, discussed by the press and parties with a stake in their improvement, and based on a regularly updated, internationally-agreed conceptual framework. This is the recipe that has made the national accounts successful; this is the way forward for all those interested in improving the measurement of inequality.

■ *We thank JEP editors Gordon Hanson and Enrico Moretti, and Timothy Taylor for detailed comments. Funding from the Center for Equitable Growth at UC Berkeley, the Sandler foundation, and the Stone foundation is gratefully acknowledged.*

References

Alvaredo, Facundo, Anthony B. Atkinson, Thomas Blanchet, Lucas Chancel, Thomas Piketty, Emmanuel Saez, Gabriel Zucman, et al. 2020. "Distributional National Accounts Guidelines: Methods and Concepts Used in World Inequality Database." WID Working Paper 2016/2, revised September 2020.

Auten, Gerald, and David Splinter. 2019. "Income Inequality in the United States: Using Tax Data to Measure Long-term Trends." http://davidsplinter.com/AutenSplinter-Tax_Data_and_Inequality.pdf.

Bricker, Jesse, Alice Henriques, Jacob Krimmel, and John Sabelhaus. 2016. "Measuring Income and Wealth at the Top Using Administrative and Survey Data." *Brookings Papers on Economic Activity* 261–331.

Bricker, Jesse, Alice Henriques Volz, and Peter Hansen. 2018. "How Much Has Wealth Concentration Grown in the United States? A Re-Examination of Data from 2001–2013." FEDS Working Paper 2018–024.

Cohen, Patricia. 2016. "Buffett Calls Trump's Bluff and Releases His Tax Data." *New York Times*, October 10. https://www.nytimes.com/2016/10/11/business/buffett-calls-trumps-bluff-and-releases-his-tax-return.html.

Colm, Gerhard, and Helen Tarasov. 1940. "Who Pays the Taxes." *Temporary National Economic Committee.*

Congressional Budget Office. 2012. *The Distribution of Household Income and Federal Taxes, 2008 and 2009.* Washington, D.C.: Congressional Budget Office.

Fixler, Dennis J., Marina Gindelsky, and David Johnson. 2020. "Measuring Inequality in the National Accounts." BEA Working Paper 2020-3.

Frémeaux, Nicolas, and Marion Leturcq. 2020. "Inequalities and the Individualization of Wealth." *Journal of Public Economics* 184: 104–45.

Institute on Taxation and Economic Policy. 2018. *Who Pays: A Distributional Analysis of the Tax Systems in All 50 States.* 6th ed. Washington DC: ITEP.

Musgrave, Richard A., John J. Carroll, Lorne D. Cook, and Lenore Frane. 1951. "Distribution of Tax Payments by Income Groups: A Case Study for 1948." *National Tax Journal* 4 (1): 1–53.

Pechman, Joseph A., and Benjamin A. Okner. 1974. *Who Bears the Tax Burden?* Washington D.C: Brookings Institution.

Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118 (1): 1–41.

Piketty, Thomas, and Gabriel Zucman. 2014. "Capital is Back: Wealth-Income Ratios in Rich Countries 1700–2010." *Quarterly Journal of Economics* 129 (3): 1255–1310.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2018. "Distributional National Accounts: Methods and Estimates for the United States." *Quarterly Journal of Economics* 133 (2): 553–609.

Saez, Emmanuel and Gabriel Zucman. 2016. "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data." *Quarterly Journal of Economics* 131 (2): 519–78.

Saez, Emmanuel, and Gabriel Zucman. 2019. "Progressive Wealth Taxation." *Brookings Papers on Economic Activity.*

Saez, Emmanuel, and Gabriel Zucman. 2019b. *The Triumph of Injustice: How the Rich Dodge Taxes and How to Make them Pay.* New York: W. W. Norton.

Saez, Emmanuel and Gabriel Zucman. 2019c. "Clarifying Distributional Tax Incidence: Who Pays Current Taxes vs. Tax Reform Analysis." Unpublished.

Saez, Emmanuel and Gabriel Zucman. 2020. "Trends in US Income and Wealth Inequality: Revising After the Revisionists." Unpublished.

Semega, Jessica, Melissa Kollar, John Creamer, and Abinash Mohanty. 2019. *Income and Poverty in the United States: 2018.* https://www.census.gov/library/publications/2019/demo/p60-266.html. (accessed August 1, 2020).

Smith, Matthew, Owen Zidar, and Eric Zwick. 2020. "Top Wealth in America: New Estimates and Implications for Taxing the Rich." Working Paper.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2019. "Capitalists in the 21st Century." *Quarterly Journal of Economics* 134 (4): 1675–1745.

Zucman, Gabriel. 2013. "The Missing Wealth of Nations: Are Europe and the U.S. net Debtors or net Creditors?" *Quarterly Journal of Economics* 128 (3): 1321–64.

Zucman, Gabriel. 2015. *The Hidden Wealth of Nations: The Scourge of Tax Havens.* Chicago: University of Chicago Press.

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 27–51

Business Incomes at the Top

Wojciech Kopczuk and Eric Zwick

When it comes to business income, the exact boundary between labor and capital can be nebulous. When a person in a partnership receives business income, in addition to regular wages, should the additional payment be categorized as "wages" or "profits"? Although this topic may seem arcane, it turns out that changes in the tax treatment of business income over time—and the resulting changes in organizational form and how business income is paid out over time—have profound implications for interpreting trends in income inequality. In addition, shifts in how business income is paid out have important consequences for interpreting tax reforms and for measuring what is counted as "labor income."

We begin with an overview of the different ways a country can choose to tax business income and how they arise in the US legal context as C-corporations, S-corporations, and partnerships. Compared with 40 years ago, a much larger share of US business income is now passed through to owner-managers rather than being subject to the corporate profits tax. We highlight the role of changing tax incentives and legal rules as crucial factors behind this shift.

Recognizing the change in how business income is being paid out and the shift to pass-through organizational forms raises questions about the measurement of top incomes, levels and trends in income and wealth inequality, and the labor and capital share of top incomes. When the rules change, the amount and timing of

■ *Wojciech Kopczuk is Professor of Economics and of International and Public Affairs, Columbia University, New York City, New York. Eric Zwick is Associate Professor of Finance, University of Chicago Booth School of Business, Chicago, Illinois. Their email addresses are wojciech.kopczuk@columbia.edu and ezwick@chicagobooth.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.4.27.

income visible on tax returns adjusts. One consequence, we argue, is that a considerable part of the increase in the top 1 percent share of income since the 1980s can be accounted for as a shift to the pass-through corporate form, not an actual rise in business income for this group. In addition, we argue that pass-through income has a substantial human capital component: for example, when the partners in a law firm or the doctors in a medical practice receive their end-of-year profit distributions, these are closer to labor income for the previous year than a return to capital.

We also discuss how changes in the form in which business income is paid out feeds into research controversies over changes in the concentration in wealth and over the progressivity of the US tax code. Income inequality in the United States has clearly risen by a variety of measures, but recent literature has been battling over the magnitude and underlying causes of increases (Piketty, Saez, and Zucman 2018; Auten and Splinter 2019). We critically discuss a number of the underlying assumptions in this work, focusing on those that pertain to business income and state which ones we currently prefer. But perhaps the key point is that conclusions here can depend heavily on underlying assumptions, which remain unsettled, but matter for fundamental questions about the interpretation and implications of changes in inequality in the United States and elsewhere. We point toward how new data might help narrow the gap between results based on competing assumptions.

At the end of the paper, we provide an overview of how these issues of pass-through and stand-alone corporate forms play out in other high-income countries, some of which offer choices for the corporate organizational form and methods of paying out business income that do not exist in the United States. We discuss fruitful avenues for future research on these topics, including the impact of relevant changes in the 2017 Tax Cuts and Jobs Act.

Pathways for Payout of Business Income

Different Forms of Businesses and their Taxation

Conceptually, there are two ways of taxing a business. One is to allocate any business income to the owners. This method is commonly used everywhere to tax small unincorporated businesses—and self-employed individuals, in particular—but it can also extend to larger firms. This approach automatically integrates the taxation of business and individual income, with the implication that business income will be taxed by the progressive marginal rates of the personal income tax. The other approach involves entity-level taxation—namely, a corporate tax. In that case, business income is taxed at the firm level and then typically taxed again when income leaves the firm and is paid to owners/shareholders.

When policymakers decide how to tax business income, they make choices about whether to have both regimes for different types of firms, or rules that decide which firms belong in each regime and how exactly personal and corporate taxes interact. Taxation on accrual—in which all profits are allocated to owners as profits are earned—implies no tax advantage to retaining funds within a firm. Conversely,

an entity-level tax with additional income taxation when owners are compensated directly may imply tax consequences of retaining earnings within a firm. In some countries, particular small businesses can be taxed in a lump-sum or withholding-tax fashion. In Australia, the corporate tax is integrated with personal income taxation so that owners receiving corporate dividends can claim offsetting credit for corporate taxes that were paid. We will discuss some international variations in a later section.

In the United States, a business more complex than a sole proprietorship can organize in multiple ways. If electing corporate form, there are two possibilities: C- or S-corporation. A C-corporation is usually the only feasible choice for publicly traded firms because it has no limit on its number of shareholders, it can have foreign or corporate owners, and it can have multiple classes of stock. Profits of a C-corporation might be distributed to shareholders as dividends (or share repurchases), but often the profits are reinvested in the firm, with the shareholders hoping to make a profit via capital gains when they sell their stock. A C-corporation falls under the traditional corporate tax regime: an entity-level tax coupled with individual-level taxes on dividends and capital gains.[1] However, firms can also finance their operations through debt, so that receipt of interest income on corporate debt becomes another possible stream of (taxable) compensation.

The S-corporation structure is more restrictive. Its shareholders are limited to individuals, estates and certain types of tax-exempt entities and trusts, and US residents. It also imposes limitations on "passive income" (which in this context includes income from royalties, dividends, interest payments, and certain other sources). In an S-corporation, profits each year must be passed through directly to the owners, which means that business income falls under the individual income tax. Most S-corporations could choose to be C-corporations, which leads to important consequences. On one hand, businesses choosing non-pass-through treatment will be subject to corporate taxation, and then shareholders will be taxed either for dividends or capital gains, but the timing and consequences of these payouts are to some extent under the owner's control. On the other hand, businesses choosing pass-through treatment avoid the corporate income tax, but each year, owners pay personal ordinary income tax treatment on their business income. With pass-through treatment, losses can also be passed through to owners, which allows for the possibility that owners could use those losses to offset other types of income. A corporation can easily switch its status between C- and S- (assuming it meets the legal conditions), except that switching can't be done more often than once every five years, and in some cases transition can entail additional one-time taxes.

An alternative to incorporation is a partnership form. Since the 1990s, as the result of state-level legal innovations, limited liability partnerships have become an option for a broader range of firms. A partnership allows for less transparency to the broad public, does not have restrictions on the types of shareholders, and allows for

[1] See Gordon and Sarada (2019) for an in-depth discussion of the role of corporate taxation.

much more flexibility in allocating income to shareholders than corporate forms do. A partnership can easily choose to be taxed as either an S- or a C-corporation through a "check-the-box" rule. A disadvantage of partnership form relative to S-corporations is that active partnership income is subject to self-employment taxation, while in the case of S-corporations the payroll tax applies only to the salary portion of income (which has to be set at a "reasonable" level).

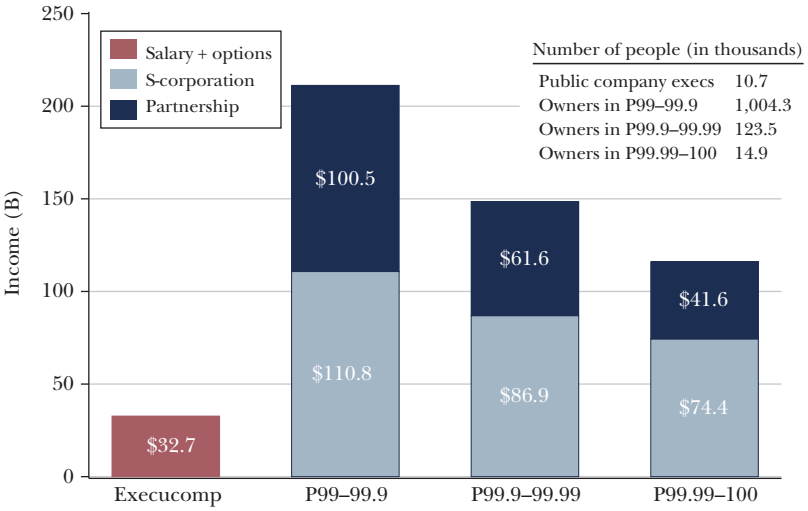
The Big Shift of Business Income to Pass-Through Firms

The category of pass-through businesses—both S-corporations and partnerships—includes, among others, consultants, lawyers, doctors, and owners of large non-publicly traded businesses, such as auto dealers and wholesale distributors. It turns out that a majority of the top income earners in the United States are owners of "pass-through" businesses (Smith et al. 2019). In 2014, 69 percent of the top 1 percent of income earners and 84 percent of the top 0.1 percent of income earners accrued some pass-through business income. In absolute terms, that amounts to more than 1.1 million pass-through owners with annual incomes above $390,000 and 140,000 pass-through owners with annual incomes of more than $1.6 million. In both number and aggregate income, these groups far surpass the top executives at public companies, who have been the focus of much inequality commentary. As shown in Figure 1, the 10,700 top public company executives earned a total of $33 billion in 2014 in salary and options. In contrast, the 14,900 business owners in the top 0.01 percent of the income distribution received more than $100 billion in income from S-corporations and partnerships. In 2014, approximately 270,000 wage earners in the top 1 percent and 27,000 wage earners in the top 0.1 percent worked for public companies, earning a total of $260 billion and $110 billion in wages and salaries, respectively. For every public company employee in the top 1 percent and top 0.1 percent, there are four and five pass-through owners, respectively.

In short, the typical top 1 percent earner is not a public company executive or tech billionaire; instead, a top earner is typically a doctor, lawyer, or the owner-operator of a middle-sized business. These top pass-through owners are predominantly working age, in contrast to the older top earners whose income comes from other categories of capital. Looking at those with more than $1 million in annual income, Smith et al. (2019) find that 60–70 percent of the millionaires who get a majority of their income from either wages or pass-through ownership are in their 40s and 50s. However, the millionaires who get a majority of their income from C-corporation dividends or other capital tend to be older, with about two-thirds falling into age brackets from their 50s to their 70s.

Since the 1986 Tax Reform Act, tax incentives have favored pass-through treatment. The trend toward more S-corporations started soon after, further encouraged by later rule changes that allowed for more shareholders. The trend toward rising partnerships followed in the mid-1990s, reflecting the state-level changes that allowed for more flexible limited liability company forms and federal guidance on how these entities would be taxed. Cooper et al. (2016) assemble data from de-identified administrative tax records on the population of US businesses linked to their

Figure 1
Pass-Through Owners Prevail at the Top of the Income Distribution

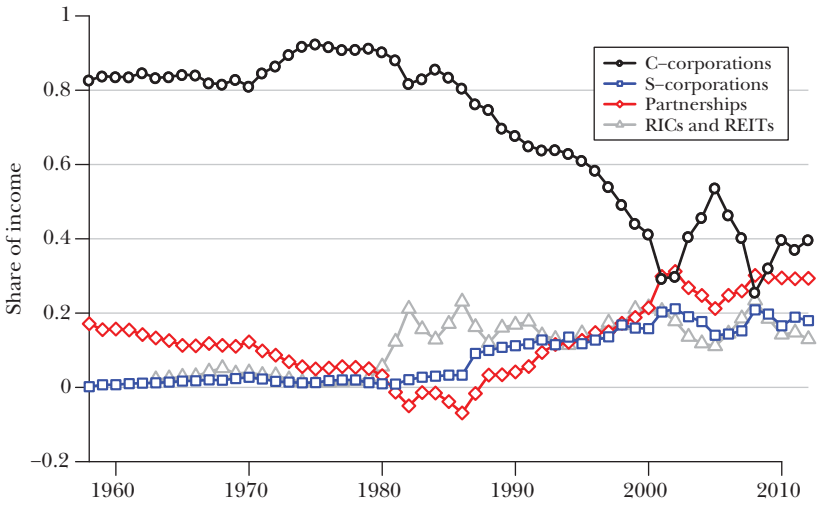


Source: Smith et al. (2019).

owners and workers and document that the role of pass-through businesses in the US economy has been rising since the Tax Reform Act of 1986. Clarke and Kopczuk (2017) also document the rise of pass-through businesses in recent decades. In 1960, the US economy had about 1 million C-corporations, 1 million partnerships, and almost no S-corporations. By 1980, the number of C-corporations had risen to 2.2 million, the number of partnerships to 1.4 million, and S-corporations had reached about 500,000. But by 2012, the number of C-corporations had declined to 1.6 million, while partnerships had climbed to 3.4 million and S-corporations to 4.2 million.

Specific C-corporations tend to have greater revenues and profits than single S-corporations and partnerships, but these shifts in the number of firms also show up clearly in business income. As shown in Figure 2, the share of business income going to C-corporations was 90 percent and higher in the late 1970s. But by 2012, C-corporations represented only about 40 percent of business income, while partnerships represented about 30 percent of business income, and S-corporations and the category of "RICs and REITs" each made up about 15 percent of business income. A "regulated investment company" (RIC), like most mutual funds or closed-end investment funds, must legally pay out at least 90 percent of its income each year to its owners; similarly, a "real estate investment trust" (REIT) must also pay a minimum of 90 percent of its income to its shareholders each year. Thus, both RICs and REITs are pass-through forms of corporate organization for certain types of mostly publicly held investment firms.

Figure 2
The Evolution of the Share of Business Income Accounted for by Different Types of Entities



Source: Clarke and Kopczuk (2017, Figure 3).

How Categorization of US Business Income Responds to Shifting Rules

When tax laws shift, and tax revenues shift in response, does the change represent a real shift in economic behavior or just a shift in accounting practices? In this section, we review the choices about how and when to receive business income. In the next section, we discuss some implications. As Gordon and Slemrod (2000) argued, shifts between receiving business income in personal or corporate form might help to explain what otherwise looks like changes in inequality or in the labor share of income.

Let's start by imagining a firm with current profits to be distributed to owners. To make the exercise concrete, we focus on a representative pass-through firm owned by a top earner using data from 2014 (as specified in Smith et al. 2019). The firm has $10 million in sales, $1 million in profits, 50 employees, and two owners. The firm makes $1.5 million in payments to the two owners: that is, it pays each owner $250,000 in salary and $500,000 in profits.

How might an enterprising owner-manager choose to receive the corporate profits? One possibility is for the owner to receive an annual bonus, which would be categorized as part of an overall wage compensation payment. Or the owner might pay herself a dividend, which would be issued based on percent ownership

of the firm. For a US pass-through entity, any profits not distributed as wages would be deemed as automatically distributed based on the owner's percent ownership of the firm, even if the money sits in the firm's bank account (this provision prevents pass-through business owners from storing funds in their firm as a method of deferring taxes). The owner would then classify these as-if distributions as ordinary business income on a personal tax return.

Each of these choices faces particular rules, but the rules do allow considerable leeway. For example, if our firm is an S-corporation and the "reasonable compensation" rules consider $250,000 an appropriate salary for such a business owner, then the owner can pay the remaining $500,000 due to her as profits—and thus avoid payroll taxes on the latter income. In contrast, a partnership would not be able to avoid the payroll tax. If the firm is a C-corporation, then the owner would likely prefer to pay this amount as a one-time bonus so that it is not included in the base for the corporate income tax.

Now consider an owner who makes a loan to the firm. In this case, the owner might receive compensation via some combination of principal returned and interest paid on that loan (subject to rules about the interest rate that can be charged). Such arrangements might make sense if the tax rate on interest payments received is lower than the tax rate on payouts, although this is not the case under the current US tax law. Alternatively, interest rates on loans against a personal residence and on direct loans against business assets need not be the same, in which case owners might arbitrage by borrowing against their personal residences rather than taking a more expensive business loan.

Finally, the owner might choose to consume through the firm. The menu of allowable fringe benefits and deductions might include meals and entertainment, club membership, travel expenses, technology, transportation, or even housing if the owner lives in the same building where the firm operates. The owner could also choose to give to charity through the firm—even a charity that the owner personally supervises—which would prevent that money from being taxed as a distribution. In administrative data, this form of income would not appear to have been paid to the owner because the firm would report these expenditures as business expenses or charitable contributions.

The consumption strategy has been less appealing recently because such deductions have become more limited in the tax rules. However, active literature in the 1970s and early 1980s considered the effect of tax preferences for non-wage compensation on the use of perks for executives (Clotfelter 1979, 1983; Long and Scott 1982; Woodbury 1983). Surveys during the Carter administration suggested that these deductions could amount to 20–30 percent of total compensation for managers at that time, and owner-managers would have a particular incentive to use this option. For example, Clotfelter (1983) focused on how tax policy affects non-wage compensation in the form of travel, meals, and entertainment expenses for sole proprietors. He references colorful contemporaneous press accounts from this time period: the president of the Philadelphia Phillies baseball team reported that "at least half the tickets were held by business [. . .which also . . .] account for

70 percent of the sales of season tickets," and a survey of compensation practices at 468 companies found that, inter alia, 53 percent paid for country club memberships and 79 percent paid for travel of spouses.[2] This line of research withered in the wake of the 1986 Tax Reform, which attenuated incentives for non-wage compensation by substantially reducing top income tax rates. The deduction for business meals and entertainment expenses was also limited in 1986 and 1993 (when deductibility was limited to 50 percent; see Schmalbeck and Soled 2009 for discussion). The 2017 tax reform repealed the business entertainment deduction altogether.

Timing issues are another major difference between the corporate tax and pass-through taxation. Pass-through taxation requires that profits be distributed to owners each year; indeed, even if business income of a pass-through firm in a given year is not actually transferred to the owner, it is subject to personal income taxation. Under the corporate tax, corporate income is taxed each year. This approach prevents business owners from completely delaying taxation while income accumulates within the firm. However, the corporation has more control over the timing of payouts to owners and/or shareholders whether via dividends or through share buybacks. In addition, an owner/shareholder of a C-corporation can defer business income from personal income taxation by investing in a firm that reinvests a substantial share of its earnings and waiting to make a profit from selling stock holdings for a capital gain at some point in the future. Of course, this choice also has to take into account the rate of return inside and outside of a business, including the fact that returns to reinvesting profits within a business can be subject to corporate taxation.

Though we usually only associate deferral of taxes with C-corporation distributions, both pass-through and traditional C-corporations can also set up retirement accounts for their owners, which allows them to defer taxation of earnings up to a certain amount until it is withdrawn. When Mitt Romney ran for President in 2012, for example, it was reported by Cohan (2012) that his private equity firm allowed partners to buy stakes in their funds with retirement account savings, which ended up earning dramatically higher returns than typical public company investments in retirement accounts.

Firms can also buy life insurance or other annuity products on behalf of owners, which has the effect of deferring income and taxes. The firm's contributions to these accounts are tax-preferred, as are the accumulated earnings on investments made by the insurance companies on behalf of their policyholders. Such arrangements are especially popular in European countries. For example, in France, the rules governing life insurance-style savings accounts ("Assurance-Vie") have relaxed over time, making such accounts the most important source of tax-deferred private savings. In recent years, these accounts provide for tax-free accumulation, occasional taxed distributions during life, and preferential inheritance tax with no

[2] Sources for these press accounts are "If Congress Taxes Those Business Perks" in *U.S. News and World Report* (February 27, 1978, pp. 53–56), and "Executives' Privileges are Under Heavy Fire but Appear Resilient" in *Wall Street Journal* (October 19, 1977, p. 1+).

contribution limits. In 2010, 83 percent of the top 1 percent of the wealth distribution had Assurance-Vie, accounting for nearly 20 percent of their total wealth (Goupille-Lebret and Infante 2018). We are not aware of research that connects the use of Assurance-Vie specifically to French business owners, though the incentives for them to participate appear strong.

Employees can also be compensated in the form of equity in the firm—either through stock grants directly or through stock options. The market value of such equity compensation is not easily observable for closely held firms, creating an opportunity to understate it (the benefit of which needs to be traded off against business-side tax consequences of reduced wage deductions).

When founders and early employees of start-up companies accept low wages in exchange for stock options, they are also deferring the recognition of income accruing to them in the form of hoped-for capital gains as their options increase in value. Only when the options are exercised, which can be many years after the shares are granted, does this income appear in administrative tax data. Naturally, for all deferred compensation—whether it be retirement accounts, life insurance, or stock options—the extent of deferral depends on the tax wedge between deferring and taking that income now, which will differ by corporate form and for wage versus non-wage income.

A final option for private firms is for the owner to plan to sell a share or all of the business—yet another method of deferring business income. Of course, a sale means losing control over the firm and gives rise to taxation of realized capital gains. Alternatively, US tax law includes a "step-up in basis" at death, which effectively forgives capital gains tax liability on assets transferred at death (for a discussion, see Kopczuk 2017). There are also estate tax-planning strategies that involve transferring some shares into a trust whose beneficiaries are the owner's children. In this case, business income accrues to the children and subsequent estate taxes or capital gains taxes can be avoided. Finally, depending on the business and type of assets, there are options for deferring or avoiding capital gains tax such as "like-kind exchanges" (especially in the case of commercial real estate) or sale to an Employee Stock Ownership Plan (ESOP).

These considerations are not theoretical: for example, evidence from different countries and times shows responses to corporate versus pass-through tax treatment. In most countries, the main choice is between sole proprietorship and a corporation. Romanov (2006) finds that incorporations in Israel respond strongly to changing tax incentives, Thoresen and Alstadsæter (2010) find similar responsiveness in Norway, Edmark and Gordon (2013) in Sweden, Sivadasan and Slemrod (2008) in India, Waseem (2018) in Pakistan, Tazhitdinova (2020) in the United Kingdom, Goolsbee (1998) in the United States using pre-World War II data, and Onji and Tang (2017) in 19th-century Japan. In the modern US economy, such responses are facilitated by the existence of pass-through business forms that allow even large businesses to be subject to individual income tax treatment. Gordon and MacKie-Mason (1994), MacKie-Mason and Gordon (1997), and Goolsbee (2004) provide evidence of the shifts between C- and S-corporation forms. Auten,

Splinter, and Nelson (2016) decompose the growth of S-corporations around the Tax Reform Act of 1986 into conversions and new incorporations, showing that conversions spiked immediately after the reform but continued through the 1990s.

Implications for Inequality

Measurement of Income Inequality

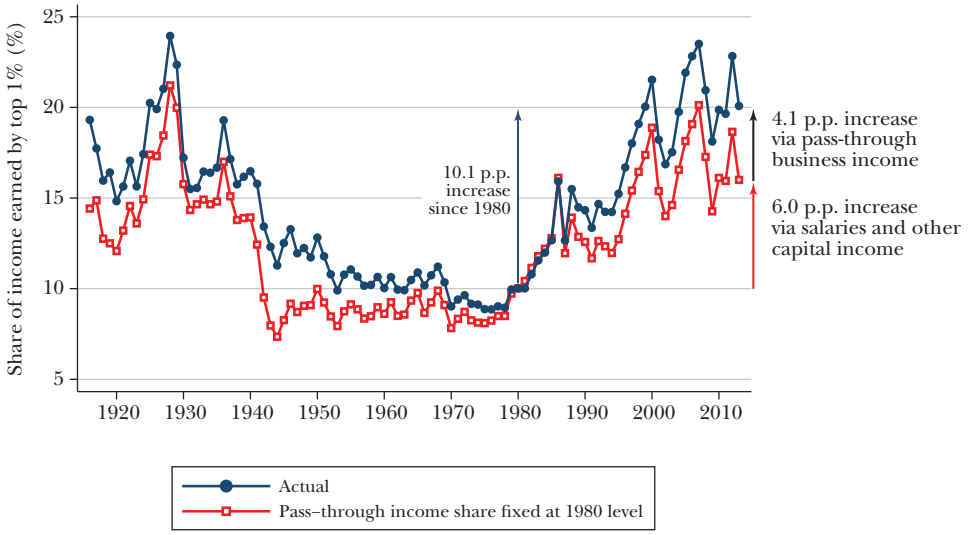
Because business income is concentrated at the top, the murky character of business income implicates several recent lines of research on income and wealth inequality. First, if business income is reported, its measurement and classification depends on whether it is reported before or after entity-level taxes and whether it takes the form of reported capital or labor. Second, there are several cases—such as consumption through the firm, retained earnings, deferred compensation, contributions to pension plans or life insurance, and other forms of tax avoidance and evasion—where such income may not be observed at all, or at least not at the individual level.

Reorganizing from C-corporation to S-corporation form can alter measures of inequality based on income tax returns. For example, after the Tax Reform Act of 1986 there was a massive conversion of C-corporations to S-corporations. The unadjusted Piketty and Saez (2003) series, which only includes fiscal income appearing on individual tax returns, shows a 4 percentage point jump in the top 1 percent share in two years, from 9 percent to 13 percent (Figure II, p. 12). This jump is certainly not a pure reflection of an underlying change in pre-tax income inequality. Instead, when firms switch from C-corporation to S-corporation form, there is a corresponding shift in income tax data from observing business income only when it is realized—and after corporate tax is paid—to observing it annually as it accrues—and before tax is paid. This induces a mechanical increase in top fiscal income shares. A number of different studies, including Piketty, Saez, and Zucman (2018) and Auten and Splinter (2019), have suggested ways to remove this bias by allocating C-corporation retained earnings and corporate taxes to individuals, but considerable controversy remains over the underlying assumptions.

The example raises the question of how to interpret trends at the top of the income distribution. In Figure 3, the line with solid dots shows the share of fiscal income received by the top 1 percent, following Piketty and Saez (2003).[3] Cooper et al. (2016) calculate what the income share of the top 1 percent would be if the share of pass-through income was held constant at its 1980 level. As shown by the hollow points in Figure 3, nearly half of the rise since 1980 in the fiscal income share of the top 1 percent comes from pass-through business. Auten and Splinter

[3]This fiscal income series includes realized capital gains, which partly obscures the 1986–88 increase in income due to corporate form reorganization. The reason is that the tax reform also raised capital gains taxes, which induced a large amount of acceleration to retime capital gains into 1986.

Figure 3
Role of Pass-Through Income in Rising Top 1 percent Income Share



Source: Cooper et al. (2016).
Note: Income in this figure is pretax fiscal income including capital gains.

(2019, see their Online Appendix Table B6) conduct a similar exercise using a panel of individual tax returns in the window around the 1986 reform and find that 20–25 percent of the short-run increase in the top 1 percent fiscal income share comes from pass-through business. Both income from new pass-throughs and limitations on allowed losses from old pass-throughs are responsible for this pattern.

Clearly, holding the share of pass-through income constant at 1980 levels does not alter the broad pattern of changes in inequality in the last century: that is, a fall in (pretax) income inequality from higher levels in the 1920s and 1930s to lower levels from the 1950s through the 1970s and then a rise in income inequality after that. However, one's perspective on the size of the rise in inequality is affected by whether one views the rise in pass-through income as an actual increase in income for those at the top of the income distribution, or whether it only means that business income that top income-earners would have received in other forms has now shifted to the pass-through channel.

In this spirit, Smith et al. (2019) use two complementary approaches to explore this question. According to their estimates, though the majority of the post-1986 growth in pass-through income reflects real economic growth, a significant share (approximately 30 percent) reflects businesses reorganizing to pass-through form without a real increase in pre-tax income inequality. This reorganization continued

through the 1990s and 2000s and even accelerated after the 2001 tax cuts during the Bush administration.

A Broader View of Inequality and the Role of Business Incomes

The income reported on tax records is roughly half of gross national income and 60 percent of net national income. Thus, Piketty, Saez, and Zucman (2018) have been pursuing "distributional national income accounts"—that is, seeking to measure a distribution of income that includes all national income. This task is intertwined with the question of business income because it involves deciding who in the income distribution should be credited with retained business income, entity-level corporate taxes, underreported business income, and pension income.

This is an area of research where, because of missing data, the assumptions play a large role, and an active debate rages on over these assumptions. Deaton (2020) summarized the current state of play in this way:

> Piketty, Saez and Zucman (PSZ) have done a great service by calculating a set of distributionally disaggregated national accounts for the United States. The basic idea is irresistible. Yet these first attempts have raised many serious difficulties that were not apparent at first. Most immediately, only about half of national income appears on individual tax returns. Allocating from tax returns is hard enough, because tax units are neither individuals nor households, but allocating the other half of national income is an immensely more difficult task, requiring assumptions that are rarely well supported by evidence, and often seem arbitrary. Because distribution is such a controversial topic, these assumptions leave plenty of scope for politically-biased challenges, in which each commentator can choose their own alternatives and get almost any result they choose, inequality is increasing, inequality is not increasing, and everything in between.

Saez and Zucman (this volume) provide a discussion of their approach. Garbinti, Goupille-Lebret, and Piketty (2018, 2020) take a similar but not identical approach for France. Auten and Splinter (2019) provide an alternative set of assumptions for the US economy. Smith et al. (2019, in their online Appendix) offer some additional comments on the methodology and explore the robustness of their results under different approaches for allocating retained earnings. Smith, Zidar, and Zwick (2020) use refined wealth estimates (described below) to improve allocation assumptions. Each of these studies proposes and defends a preferred inequality series.

Fundamentally, though, as highlighted by Deaton's comment, there is currently no information allowing us to assign certain macroeconomic aggregates from the national income and product accounts to individuals, and the lack of micro data means that researchers resort to imputations. The vast differences in the resulting levels of inequality and trends reflect the extent of underlying uncertainty. That uncertainty is not explicit when one presents the results as a definitive series

measuring inequality rather than as estimates relying on a large number of assumptions and therefore with a large margin of error. For that reason, we'll refer to these results as "imputations" rather than "estimates."

Much business income is not directly assigned to individuals in tax data, and one possible approach is to make assumptions about asset ownership across the income distribution. Piketty, Saez, and Zucman (2018) start with the income tax data showing taxable capital income received from financial investments or other assets and then, with a set of auxiliary assumptions, infer the underlying distribution of wealth. The wealth imputation method proposed by Saez and Zucman (2016) scales up, or "capitalizes," the income observed on tax returns to impute wealth. Piketty, Saez, and Zucman (2018, following Saez and Zucman 2016) infer this distribution in broad asset categories: for example, fixed income, stocks, pass-through business, housing, and pensions. For example, if the tax data reveals a certain level of interest payments received, a researcher can then try to infer what wealth was needed to receive these interest payments. Clearly, this approach relies upon having an accurate mapping of income to wealth, or equivalently, knowing the rates of return earned on different types of income by different groups of people. Saez and Zucman (2016) deploy the simplifying assumption that all tax units get the same rate of return within an asset class; Piketty, Saez, and Zucman (2018) also assume the same rate of return across asset classes, but instead use a "divide-by-two" method to attribute wealth to individuals within married tax units. The tax return data and wealth imputations are then combined with aggregate data from the national income accounts to impute distributional national income accounts. The overall approach is somewhat circular—we go from (observed) income to (unobserved) wealth to (unobserved) income—but the results by construction will "add up" to published aggregates.

If one is interested in the distribution of household wealth at a point in time, income tax data is not the natural starting point. The natural alternative is the Survey of Consumer Finances done triennially by the Federal Reserve (Bricker et al. 2016). It has its weaknesses—including lack of coverage of the extreme top of the wealth distribution, and the modern version of the survey only goes back to 1983 (with precursor surveys going further back, but more consistent design since 1989)—but it does not require imputation exercises. It also allows for observing the joint distribution of income and wealth, avoiding the need for another set of assumptions. The Federal Reserve now builds on the Survey of Current Finances to construct the Distributional Financial Accounts (Batty et al. 2019) that provide quarterly estimates of the US household income and wealth distribution since 1989.

Other assumptions need to be made in moving from the income data to estimates of the underlying wealth that generates business income: for example, how to attribute the ownership of C-corporations. Saez and Zucman (2016) assume that C-corporation wealth directly held by households is distributed in proportion to the sum of dividends and realized capital gains. Smith, Zidar, and Zwick (2020) propose an alternative assumption that weighs dividends and realized gains based on their relative informativeness in predicting stock wealth in the Survey of

Consumer Finances; this method's results give less weight to realized capital gains. For pass-through business, Saez and Zucman (2016) assume that pass-through business income has equal returns across industries. Alternatively, Smith, Zidar, and Zwick (2020) adjust the valuation of pass-through business for the human capital component of business income and allow returns to vary across narrowly defined industries. Another notable assumption is whether receiving pension income means that you "own" a certain amount of pension wealth. In a defined contribution plan, the pension recipient does own underlying wealth (although the amount is not directly observed in administrative data); in a defined benefit plan, the recipient of pension income does not hold such wealth—and if the plan is underfunded, the underlying wealth may not exist. Both Saez and Zucman (2016) and Smith, Zidar, and Zwick (2020) use a combination of wages and pension distributions to infer pension wealth (see also Sabelhaus and Henriques Volz 2020). Both papers do not attribute any wealth based on the "off-balance-sheet" public pension benefit known as Social Security (Catherine, Miller, and Sarin 2020; Sabelhaus and Henriques Volz 2020).

Distributional national income accounts require allocation of income from assets not reported on tax returns, such as C-corporations, pensions, and underreported business income. Thus, the imputations for wealth inequality then feed back into the imputation of distributional national income accounts, along with tax data and macro data from the national income accounts. In this way, constructed measures of inequality of wealth are used to impute the distributional national income accounts for top income shares (and shares for labor and capital income shares, as discussed in the next section).

Several studies have raised concerns about the set of assumptions required for this imputation. For example, one concern is that the equal returns assumption can bias wealth estimates toward the top when top wealthholders actually earn higher returns than average. Kopczuk (2015) suggests these adjustments are especially important when average returns are close to zero, such as was the case for interest rates in the wake of the Great Recession because a relatively small shift between two very low rates of return will imply a large shift in implied wealth (say, from 0.5 percent to 1.0 percent cuts the implied wealth by a factor of 2). Fagereng et al. (2016), Bricker, Henriques Volz, and Hansen (2019), and others also emphasize the evidence that those at the top of the income distribution typically get higher returns for a given asset class.

Smith, Zidar, and Zwick's (2020) preferred results, using a range of literature to estimate rates of return across wealth groups and geographic areas, find a rise in wealth concentration similar to the data of the Survey of Consumer Finance. The assumption of equal returns across asset classes also implies that fixed income wealth should be rapidly increasing as a share of top portfolios. In contrast, the evidence-based patterns of returns they use lead to an estimated portfolio concentration of top wealth holders that aligns reasonably well with estimates from the Survey of Consumer Finance and from estate tax data, in that private business is more important than fixed income and rivals or exceeds public equity holdings at

the top. Smith, Zidar, and Zwick (2019) also use their wealth results to construct distributional income estimates, which allocate components of capital income not observed on tax returns: for example, fixed income earned in non-taxable accounts, retained earnings of C-corporations, accumulated returns to assets held in pension accounts, and taxes whose statutory incidence does not fall on individual owners. They find top income shares somewhat lower than in Piketty, Saez, and Zucman (2018), but the trends in concentration are similar. However, the composition of top incomes and their recent growth skews much more toward labor than in the Piketty, Saez, and Zucman (2018) series.

Other important assumptions concern allocating "underreported" income (that is, income not reported to tax authorities) and pension income. Auten and Splinter (2019) propose and defend alternative assumptions for these categories. The largest disagreement they emphasize concerns how to allocate underreported income for non-corporate business. Auten and Splinter (2019) use IRS audit data to allocate underreported income; Piketty, Saez, and Zucman (2018) allocate this income in proportion to positive fiscal non-corporate business income, under the assumption that the distributions of observed and unobserved income in this category are the same. The fact that assumptions about underreported income are so consequential highlights the central role missing data on business income plays in controversies about income inequality.[4]

In order to allocate all of national income post-tax (rather than pre-tax), additional assumptions are needed that are not necessarily related to wealth, such as who benefits from defense spending and other public goods. Piketty, Saez, and Zucman (2018) allocate these public goods "neutrally" in proportion to income; in contrast, Auten and Splinter (2019) argue that a significant proportion of such spending should be allocated equally across people.

Finally, Piketty, Saez, and Zucman (2018) take an additional step in using the pretax distributional national income accounts together with aggregate tax payments—such as payroll tax, sales tax, property tax, estate tax, and corporate tax—to estimate the post-tax distribution of income and thereby a measure of broad tax rate progressivity. Conceptually, this measure of tax rates has all of an income group's imputed national income in the denominator and all of their imputed tax payments in the numerator. Saez and Zucman (2019) take a similar approach but make different assumptions to measure tax rate progressivity. Again, a number of assumptions undergird such calculations. A main focus of our discussion has been the flexibility in allocating business income across various corporate forms, and a follow-up question that arises here is the incidence of the corporate tax across income groups. Another one is the incidence of the payroll tax. Questions less related to the allocation of business income include the incidence of the sales tax and the treatment of various social support programs that include transfers and refundable tax credits, such as the earned income tax credit.

[4]Sabelhaus and Park (2020) also note the particularly large gap between the national income and product account and the Survey of Consumer Finances for non-corporate business incomes.

As Splinter (2019, 2020) points out, the assumptions in Saez and Zucman (2019) lead to a conclusion that the overall US tax code is more-or-less proportional. In contrast, a wide variety of other sources including the Joint Committee on Taxation, the US Treasury, the Congressional Budget Office, the Urban-Brookings Tax Policy Center, and Piketty and Saez (2007) all find that the federal tax system is progressive, although somewhat less so than it used to be. The assumptions in Saez and Zucman (2019) are often non-standard and a departure from the widely accepted practice by agencies and economic literature, including their own work (Piketty, Saez, and Zucman 2018). At the top, they make an unusual "statutory incidence" assumption to load the full corporate tax burden on shareholders (rather than allocating part of it to other capital or labor), although they continue the standard practice of disregarding statutory incidence and assigning the burden of sales taxation to consumers and the employer portion of payroll tax to workers—even though these taxes are also legally and administratively collected from firms. They also make unorthodox assumptions about the distinction between taxes and transfers and assign the sales tax burden based on transfers-financed consumption, while not including transfers in measures of income, thereby artificially inflating effective tax rates at the bottom of the distribution (for details, see Splinter 2019; Kopczuk 2019).

Taking Stock

This task of developing distributional national income accounts that cover all of national income is clearly an active area of research.[5] We see value in continuing attempts to reconcile these different approaches to estimating wealth, imputing all of national income to different groups, and thinking through the tax incidence and tax burden issues. Given the current state of this research, it would seem appropriate here though to acknowledge the vast uncertainty of any imputations in a much more systematic way than has been the case so far.

Yet another challenge is the changing tax treatment of various categories of business income, which makes comparisons across years very challenging. First, the tax treatment of capital gains changes over time, which affects imputed stock wealth of C-corporations and imputed retained earnings. Second, the tax incentives to shelter income in corporations or through corporate consumption changes over time, which affects how much income we observe on tax returns. Third, stock options appear partly as wages and partly as capital gains (when realized), which clouds both the timing and reported nature of this important component of top executive compensation. Fourth, the tax treatment and rules for pensions have changed over time, which can affect the amount of business income distributed into pension savings. Finally, the tax treatment of business losses means that some

[5]For other attempts to allocate income, transfers, and taxes not observed on individual tax returns or in household surveys, see the work from the Congressional Budget Office (for example, Congressional Budget Office 2016) and from economists at the US Bureau of Economic Analysis, including Fixler, Gindelsky, and Johnson (2019) and works cited therein.

wealthy individuals can appear to be at the bottom of the income distribution in a given year despite having substantial cash incomes, and this issue has also changed over time.[6]

Future data collection and refinements of methodology could address these various controversies. First, if partnerships and C-corporations were required to trace and report their ultimate owners, this linked data could be used to allocate macroeconomic business income, in the spirit of Cooper et al. (2016). Second, Internal Revenue Service data from random audits could be used to improve our understanding of underreported income, refine inequality estimates, and reconcile disputes. For example, DeBacker et al. (2020) use random audit data from 2006 to 2014 and find that because top earners have higher rates of compliance, measures of income inequality are lower after accounting for noncompliance. Third, more data collection on retirement account balances and portfolio composition could help allocate the assets and income flows accruing in these accounts.

Labor versus Capital Income

Researchers care about the allocation of "labor income" and "capital income" for at least three reasons. First, it provides insight into the role of technology and economic factors versus institutions and public policy in driving economic inequality. Second, it speaks to the nature of typical paths to the top of the income distribution and thus offers insights about intergenerational mobility and barriers to such mobility. Third, studying the labor share can guide policy reforms designed to reduce inefficiencies in markets, alter the post-tax distribution of income, and raise tax revenues.

For all the reasons given in the discussion above, when we wish to compare labor to capital income—especially over time or across countries—we must take into account the effects of changes in the tax code on how income is categorized. Smith et al. (2019) present a comprehensive analysis of pass-through business income with the goal of answering the question: how important is human capital at the top of the US income distribution? Human capital in this research is defined broadly to refer to all factors embodied in people, including labor supply, networks, reputation, and rent-seeking ability. Human capital contrasts with nonhuman or financial capital because (in the modern economy) human capital can't be sold, and it is not bequeathed at death. Combining descriptive analysis with natural experiments, Smith et al. (2019) find that human capital, as opposed to financial capital, remains central to rising top incomes in the US economy.

[6]In their imputations, Auten and Splinter (2019) attempt to account for business losses in three cases: 1) adding net operating loss carryovers from past years because they are unrelated to current national income; 2) applying the limit on business losses from the Tax Reform Act of 1986 to data from before the passage of that law; and 3) allocating underreported income following the audit data analysis of Johns and Slemrod (2010). In contrast, Piketty, Saez, and Zucman (2018) only use positive business profits to impute wealth and business income.

This finding depends crucially on how we think about pass-through income, which Smith et al. (2019) estimate to have a human capital share of 75 percent even though it appears for tax purposes as business profits. They construct this estimate by following firms after premature owner deaths and retirements and observing the impact of withdrawing owners from their firms. When ignoring pass-through income, it appears that only a minority of top earners are human-capital rich. However, when defining labor income comprehensively to include that share of pass-through income, this assessment reverses: most top earners are human-capital rich, not financial-capital rich, as shown in Figure 4. In follow-on work, Smith et al. (2020) find that neglecting how taxes influence income reporting would lead us to overstate how much economic growth has accrued to capital instead of labor since the 1980s. Thus, they add yet another factor that can help account for the recent decline in the labor share of national income in the US economy (Elsby, Hobijn, and Şahin 2013; Karabarbounis and Neiman 2014; Autor et al. 2020; de Loecker, Eeckhout, and Unger 2020).

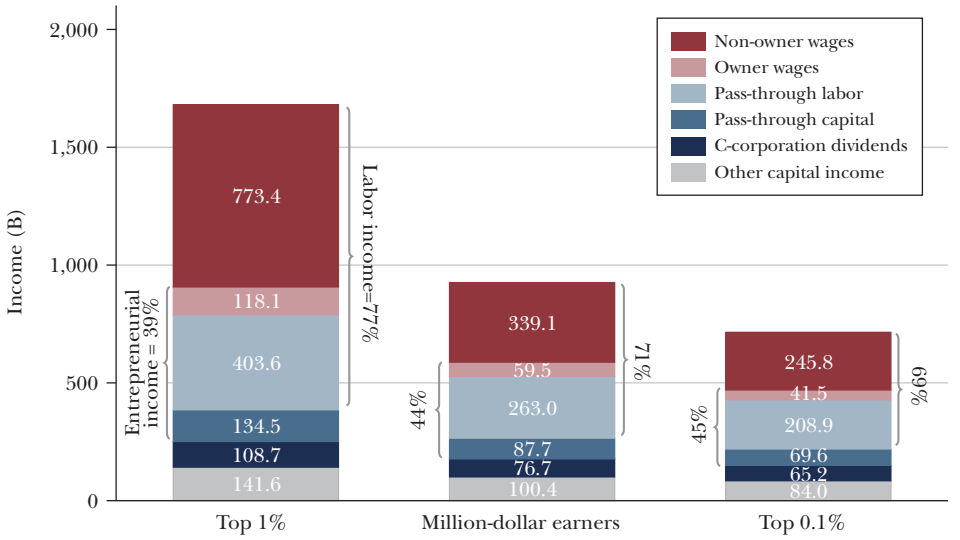
Again, a unifying message is that the underlying assumptions—especially those relevant to how business income is treated and wealth is estimated—will strongly affect one's view of the role of labor and capital income. In Piketty, Saez and Zucman (2018), the estimate of rapidly growing wealth underlies the finding that top capital shares have surged in the past 20 years, reaching 56 percent in 2014. Conversely, the alternative assumptions in Smith, Zidar, and Zwick (2019) imply that, in 2014, only 41 percent of income for the top 1 percent comes from capital. Approximately half of this adjustment comes from differences in wealth estimates discussed earlier. The remainder arises because of the Smith et al. (2019) allocation of 75 percent of pass-through income to labor, rather than attributing it all to capital.

To be clear, our reading of the evidence based on our preferred assumptions is not that inequality in America is low or that it has not increased. Rather our reading is that the increase has been more modest than some well-known estimates suggest. In addition, we believe that the nature of that increase—what factors contribute, who benefits—skews away from the passive capital highlighted in Piketty (2014) and toward human capital, labor, and entrepreneurial activity. We stress also that this belief does not imply the returns to human capital at the top are fair, nor that they necessarily reflect the social returns to labor, rather than the private returns, which could well include unproductive or even destructive activity (Baumol 1990; Murphy, Shleifer, and Vishny 1991).

Some International Perspective and Comparisons

With the US shift to more widespread use of pass-through taxation of business income, the United States now taxes business income quite differently from some other countries. The US economy now taxes about 40 percent of business income at the corporate or entity level, while for the United Kingdom, Canada, and Australia during the last 30 years, 65–80 percent of the business income that is reported on

Figure 4
Are Top Earners Human-Capital Rich?



Source: Smith et al. (2019). Shares of Fiscal Income.

tax returns is subject to entity-level or corporate taxation, rather than pass-through taxation (Clarke and Kopczuk 2017, see Fig. 1). However, the US economy is not an isolated exception. The Joint Committee on Taxation (2013) reports that in 2007, only 34 percent of business incomes in Germany were subject to corporate tax and the corresponding number for Japan was 50 percent.

The rules that guide pass-through taxation of business income vary by country. Sole proprietors are usually taxed by individual income tax or, sometimes, through alternative small business tax regimes. Corporate tax treatments apply to large firms. In between, there are usually some lines drawn concerning limited liability and organizational form. US tax law does not tie pass-through treatment to a lack of limited liability: instead, pass-through of business income in US law applies not just to sole proprietors and farm income, but also to some incorporated businesses (S-corporations) and partnerships. A similar approach is also used in Canada, Germany, and the United Kingdom, with partnerships generally eligible for pass-through treatment (even if they have limited liability), but these countries have no equivalent to S-corporations. Australia taxes most partnerships as companies, as long as at least one partner is subject to limited liability (Joint Committee on Taxation 2013). As another example, Poland nominally ties pass-through treatment to lack of limited liability, but allows a hybrid form with both limited and unlimited liability partners to be eligible as well.

On the international stage, comparisons of top income shares and labor/capital shares ultimately derive to a large extent from tax data. While researchers and statistical agencies do attempt to adjust for some of the measurement issues discussed here, a systematic analysis of the implications for international comparisons remains to be done. For researchers, problems arise in both the measurement of retained business income and in attempts to attribute that income to specific individuals. Gollin (2002) provided an early demonstration that correcting for different treatments of self-employment can reconcile large cross-country differences in factor income shares. More recently, Gutiérrez and Piton (forthcoming) argue that, after correcting for inconsistent treatment of entrepreneurial income (and the inclusion of housing rents in the corporate sector), the decline of the labor share is no longer apparent in advanced economies outside the United States and Canada.

It seems likely that the issues discussed in this paper can make a large difference in other countries too. For example, in many European countries, such as in France where income inequality series based on tax data often imply low and stable inequality, we know that closely held private businesses are even more important for economic activity than in the United States. These countries often have tax rules that encourage business owners to keep income within the firm and off their personal tax returns. In Norway, Alstadsæter et al. (2016) show that omitting retained business income leads to a large mismeasurement of inequality; conversely, accounting for it doubles the income share of the top 1 percent and triples the share of the top 0.1 percent in some years. They find that in the Norwegian data, these issues also affect the trends in inequality in the aftermath of a reform that created strong incentives for businesses to retain earnings. Alstadsæter, Kopczuk, and Telle (2014) find some evidence that retained business earnings were disproportionately invested in financial instruments and durable goods (cars, ships, planes) and thus may have substituted for private investment or consumption. Atkinson (2007) estimates that during the 1950s and early 1960s in the United Kingdom, including retained company profits raises income shares of the top 1 percent (excluding capital gains) by about half. Burkhauser, Hahn, and Wilkins (2015) show that a 1985 Australian tax reform captured a larger share of capital gains and corporate profits on individual tax returns, thereby increasing measured income shares of the top 1 percent by about one-sixth.

As another example of potential issues that arise, return to the role of life insurance in France mentioned earlier. Garbinti, Goupille-Lebret, and Piketty (2018) suggest that retained earnings and corporate dividends in France were each around 10–12 percent of GDP circa 2014. Moreover, dividends paid by French firms as a share of GDP have roughly doubled since 1990. This rise coincided with the expanding importance of life insurance assets (Assurance-Vie), which contain large amounts of indirectly held corporate equity for overall national wealth in France (Piketty 2011; Garbinti, Goupille-Lebret, and Piketty 2020). In the US data, retained earnings and dividends are each only about 4–5 percent of GDP during this period, consistent with a larger role for pass-through firms in the United States. Clearly, how a researcher decides to measure and attribute total

business income and retained earnings can influence measures of inequality and the labor share of income.

Attributing business income to individual owners is complex in other countries, as well, although some countries allow for linking individual tax information to business ownership and accounting data. Examples include work in Denmark (le Maire and Schjerning 2013), Canada (Wolfson et al. 2016), Chile (Fairfield and Jorratt De Luis 2016), Norway (Alstadsæter et al. 2016), and Finland (Harju and Matikka 2016), each of which uses direct links between firms and owners to correct for unobserved, unrealized income.

It would be a useful research project to make a systematic comparison of the rules regarding taxation of business income across countries. Such a project requires thinking about different organizational forms and their flexibility, the role of limited liability, and tax incentives associated with both corporate and non-corporate treatment. In turn, the different approaches to realizing business income have implications for how and when business income is reported and taxed, which in turn, has consequences for data availability.

Looking Forward

Business income reflects a mix of capital and labor income. The implications of this fact require a nuanced understanding of business activity and a thorough understanding of the various connections amongst payout, retained earnings, corporate and non-corporate profits, employee compensation, and the compensation of owner-managers. We believe that a bottom-up, micro-based approach to these questions is most likely to be productive.

We see a number of exciting research directions related to incentives and business incomes at the top. First, as the complexities of the Tax Cuts and Jobs Act of 2017 unfold, a number of research opportunities should emerge. On one hand, the law reduced the marginal and effective corporate tax rates, creating for the first time since 1987 a stronger incentive to shift business income away from a pass-through to a C-corporation structure. As an offset for capital-intensive pass-throughs that are more likely to consider C-corporation form in the first place, the 2017 legislation also introduced a new tax deduction ("Section 199A deduction") on personal income tax returns that amounts to a 20 percent reduction in taxes on business income in this form. As an acknowledgement of the incentives to characterize entrepreneurial income in the tax-preferred form, this rate is not available to a large number of "specified service businesses," including lawyers, doctors, consultants, and similar types of firms that rely primarily on human capital. Goodman et al. (2019) simulate the effect of the 199A deduction for pass-through owners based on 2016 data and conclude that while it benefits business owners throughout the income distribution, over 72 percent of tax savings accrues to the top 5 percent. Henry, Plesko, and Utke (2018) discuss the complex interaction of tax incentives regarding the choice of organizational form in the aftermath of the 2017 legislation.

Second, there is much work to be done in countries outside the United States in drawing links from private businesses to their owners and studying the implications for inequality and tax policy. In addition to the papers already mentioned, Miller, Pope, and Smith (2019) and Aghion et al. (2019), who use newly assembled data on the United Kingdom and France, respectively, are prominent recent examples. We have much to learn from how different incentive structures and rules in other countries affect the measurement and realization of top business income. For example, we are not aware of research that has connected the large pension and insurance industries outside the United States to trends in top income shares and the income-realization behavior of owner-managers.

Finally, future changes in the rates of corporate, personal, or capital gains taxation will further alter the balance between different organizational forms. Steps to increase transparency of gains to wealth are likely to have differential effects across corporate forms as well. For example, valuation of assets for the purposes of a wealth tax is straightforward for publicly traded firms, but much less so for closely held firms. Thus, a wealth tax, or other steps like requiring financial assets to be marked-to-market each would tend to make ownership forms with less effective transparency, like partnerships and S-corporations, more appealing. Both the public finance literature in particular and, more broadly, any study relying on administrative tax data should be aware of the need to take shifts in organizational form of businesses into account.

■ *We benefited from helpful comments from Timothy Taylor, Enrico Moretti, Gordon Hanson, Alan Auerbach, Jerry Auten, Roger Gordon, Daniel Reck, John Sabelhaus, David Splinter, and Owen Zidar.*

References

Aghion, Philippe, Vlad Ciornohuz, Maxime Gravoueille, and Stefanie Stantcheva. 2019. "Reforms and Dynamics of Income Evidence Using New Panel Data." Unpublished.

Alstadsæter, Annette, Wojciech Kopczuk, and Kjetil Telle. 2014. "Are Closely Held Firms Tax Shelters?" *Tax Policy and the Economy* 28 (1): 1–32.

Alstadsæter, Annette, Martin Jacob, Wojciech Kopczuk, and Kjetil Telle. 2016. "Accounting for Business Income in Measuring Top Income Shares: Integrated Accrual Approach Using Individual and Firm Data from Norway." NBER Working Paper 22888.

Atkinson, Anthony B. 2007. "The Distribution of Top Incomes in the United Kingdom 1908–2000." In *Top Incomes over the Twentieth Century: A Contrast between Continental European and English-Speaking Countries*, edited by Anthony B. Atkinson and Thomas Piketty, 82–140. Oxford: Oxford University Press.

Auten, Gerald, and David Splinter. 2019. "Income Inequality in the United States: Using Tax Data to

Measure Long-Term Trends." Unpublished.

Auten, Gerald, David Splinter, and Susan Nelson. 2016. "Reactions of High-Income Taxpayers to Major Tax Legislation." *National Tax Journal* 69 (4): 935–64.

Autor, David, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." *Quarterly Journal of Economics* 135 (2): 645–709.

Batty, Michael, Jesse Bricker, Joseph Briggs, Elizabeth Holmquist, Susan Hume McIntosh, Kevin B. Moore, Eric Reed Nielsen, et al. 2019. "Introducing the Distributional Financial Accounts of the United States." Finance and Economics Discussion Series Working Paper 2019-017.

Baumol, William J. 1990. "Entrepreneurship: Productive, Unproductive, and Destructive." *Journal of Political Economy* 98 (5): 893–921.

Bricker, Jesse, Alice Henriques Volz, and Peter Hansen. 2019. "Wealth Concentration in the U.S. after Augmenting the Upper Tail of the Survey of Consumer Finances." *Economics Letters* 184.

Bricker Jesse, Alice Henriques Volz, Jacob Krimmel, and John Sabelhaus. 2016. "Measuring Income and Wealth at the Top Using Administrative and Survey Data." *Brookings Papers on Economic Activity* 261–331.

Burkhauser, Richard V., Markus H. Hahn, and Roger Wilkins. 2015. "Measuring Top Income Using Tax Record Data: A Cautionary Tale from Australia." *Journal of Economic Inequality* 13: 181–205.

Catherine, Sylvain, Max Miller, and Natasha Sarin. 2020. "Social Security and Trends in Inequality." Unpublished.

Clarke, Conor, and Wojciech Kopczuk. 2017. "Business Income and Business Taxation in the United States since the 1950s." *Tax Policy and the Economy* 31: 121–59.

Clotfelter, Charles T. 1979. "Equity, Efficiency, and the Tax Treatment of In-Kind Compensation." *National Tax Journal* 32 (1): 51–60.

Clotfelter, Charles T. 1983. "Tax-Induced Distortions and the Business-Pleasure Borderline: The Case of Travel and Entertainment." *American Economic Review* 73 (5): 1053–65.

Cohan, William. 2012. "What's Really Going on with Mitt Romney's $102 Million IRA." *The Atlantic*, September 10. https://www.theatlantic.com/politics/archive/2012/09/whats-really-going-on-with-mitt-romneys-102-million-ira/261500/.

Congressional Budget Office. 2016. *The Distribution of Household Income, 2016*. Washington, D.C.: Congressional Budget Office.

Cooper, Michael, John McClelland, James Pearce, Richard Prisinzano, Joseph Sullivan, Danny Yagan, Owen Zidar, and Eric Zwick. 2016. "Business in the United States: Who Owns it and How Much Tax Do They Pay?" *Tax Policy and the Economy* 30 (1): 91–128.

Deaton, Angus. 2020. "Beyond GDP." *Survey of Current Business*.

DeBacker, Jason, Bradley Heim, Anh Tran, and Alexander Yuskavage. 2020. "Tax Noncompliance and Measures of Income Inequality." *Tax Notes* February: 1103–18.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2020. "The Rise of Market Power and the Macroeconomic Implications." *Quarterly Journal of Economics* 135 (2): 561–644.

Edmark, Karin, and Roger H. Gordon. 2013. "The Choice of Organizational Form by Closely Held Firms in Sweden: Tax versus Non-Tax Determinants." *Industrial and Corporate Change* 22 (1): 219–43.

Elsby, Michael W. L., Bart Hobijn, and Ayşegül Şahin. 2013. "The Decline of the U.S. Labor Share." *Brookings Papers on Economic Activity* 2013: 1–63.

Fagereng, Andreas, Luigi Guiso, Davide Malacrino, and Luigi Pistaferri. 2016. "Heterogeneity in Returns to Wealth and the Measurement of Wealth Inequality." *American Economic Review* 106 (5): 651–55.

Fairfield, Tasha, and Michel Jorratt De Luis. 2016. "Top Income Shares, Business Profits, and Effective Tax Rates in Contemporary Chile." *Review of Income and Wealth* 62 (S): S120–44.

Fixler, Dennis, Marina Gindelsky, and David Johnson. 2019. "Improving the Measure of the Distribution of Personal Income." *AEA Papers and Proceedings* 109: 302–06.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty. 2018. "Income Inequality in France, 1900–2014: Evidence from Distributional National Accounts (DINA)." *Journal of Public Economics* 162: 63–77.

Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty. 2020. "Accounting for Wealth-Inequality Dynamics: Methods, Estimates, and Simulations for France." *Journal of the European Economic Association*.

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

Goodman, Lucas, Katherine Lim, Bruce Sacerdote, and Andrew Whitten. 2019. "Simulating the 199A Deduction for Pass-through Owners." US Department of Treasury Office of Tax Analysis Working

Paper 118.
Goolsbee, Austan. 1998. "Taxes, Organizational Form, and the Deadweight Loss of the Corporate Income Tax." *Journal of Public Economics* 69 (1): 143–52.
Goolsbee, Austan. 2004. "The Impact of the Corporate Income Tax: Evidence from State Organizational Form Data." *Journal of Public Economics* 88 (11): 2283–99.
Gordon, Roger H., and Jeffrey K. MacKie-Mason. 1994. "Tax Distortions to the Choice of Organizational Form." *Journal of Public Economics* 55 (2): 279–306.
Gordon, Roger, and Sarada. 2019. "The Role of the Corporate Tax." *Cambridge Elements in Public Economics*.
Gordon, Roger H., and Joel B. Slemrod. 2000. "Are 'Real' Responses to Taxes Simply Income Shifting Between Corporate and Personal Tax Bases?" In *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, edited by Joel B. Slemrod, 240–80. New York: Russell Sage Foundation.
Goupille-Lebret, Jonathan, and Jose Infante. 2018. "Behavioral Responses to Inheritance Tax: Evidence from Notches in France." *Journal of Public Economics* 168: 21–34.
Gutiérrez, Germán, and Sophie Piton. Forthcoming. "Revisiting the Global Decline of the (Non-Housing) Labor Share." *American Economic Review: Insights*.
Harju, Jarkko, and Tuomas Matikka. 2016. "The Elasticity of Taxable Income and Income-Shifting: What Is 'Real' and What Is Not?" *International Tax and Public Finance* 23: 640–69.
Henry, Erin, George A. Plesko, and Steven Utke. 2018. "Tax Policy and Organizational Form: Assessing the Effects of the Tax Cuts and Jobs Act of 2017." *National Tax Journal* 71 (4): 635–60.
Johns, Andrew, and Joel Slemrod. 2010. "The Distribution of Income Tax Noncompliance." *National Tax Journal* 63 (3): 397–418.
Joint Committee on Taxation. 2013. *Foreign Passthrough Entity Use in Five Selected Countries*. Washington, DC: Staff of the Joint Committee on Taxation.
Karabarbounis, Loukas, and Brent Neiman. 2014. "The Global Decline of the Labor Share." *Quarterly Journal of Economics* 129 (1): 61–103.
Kopczuk, Wojciech. 2015. "What Do We Know about the Evolution of Top Wealth Shares in the United States?" *Journal of Economic Perspectives* 29 (1): 47–66.
Kopczuk, Wojciech. 2017. "U.S. Capital Gains and Estate Taxation: A Status Report and Directions for a Reform." In *The Economics of Tax Policy*, edited by Alan Auerbach and Kent Smetters, 265–91. Oxford: Oxford University Press.
Kopczuk, Wojciech. 2019. "Comment on Saez and Zucman's 'Progressive Wealth Taxation." *Brookings Papers on Economic Activity* Fall (2019): 1–13.
le Maire, Daniel, and Bertel Schjerning. 2013. "Tax bunching, Income Shifting and Self-Employment." *Journal of Public Economics* 107: 1–18.
Long, James E., and Frank A. Scott. 1982. "The Income Tax and Nonwage Compensation." *Review of Economics and Statistics* 64 (2): 211–19.
MacKie-Mason, Jeffrey K., and Roger H. Gordon. 1997. "How Much Do Taxes Discourage Incorporation?" *Journal of Finance* 52 (2): 477–506.
Miller, Helen, Thomas Pope, and Kate Smith. 2019. "Intertemporal Income Shifting and the Taxation of Owner-Managed Businesses." Institute for Fiscal Studies Working Paper W19/25.
Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny. 1991. "The Allocation of Talent: Implications for Growth." *Quarterly Journal of Economics* 106 (2): 503–30.
Onji, Kazuki, and John P. Tang. 2017. "Taxes and the Choice of Organizational Form in Late Nineteenth Century Japan." *Journal of Economic History* 77 (2): 440–72.
Piketty, Thomas. 2011. "On the Long-Run Evolution of Inheritance: France 1820–2050." *Quarterly Journal of Economics* 126 (3): 1071–1131.
Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press.
Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118 (1): 1–41.
Piketty, Thomas, and Emmanuel Saez. 2007. "How Progressive Is the U.S. Federal Tax System? A Historical and International Perspective." *Journal of Economic Perspectives* 21 (1): 3–24.
Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2018. "Distributional National Accounts: Methods and Estimates for the United States." *The Quarterly Journal of Economics* 133 (2): 553–609.
Romanov, Dmitri. 2006. "The Corporation as a Tax Shelter: Evidence from Recent Israeli Tax Changes." *Journal of Public Economics* 90 (10–11): 1939–54.
Sabelhaus, John, and Alice Henriques Volz. 2020. "Social Security Wealth, Inequality, and Lifecycle

Saving." NBER Working Paper 27110.

Sabelhaus, John, and Somin Park. 2020. *U.S. Income Inequality Is Worse and Rising Faster than Policymakers Probably Realize*. Washington DC: Washington Center for Equitable Growth.

Saez, Emmanuel, and Gabriel Zucman. 2016. "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data." *Quarterly Journal of Economics* 131 (2): 519–78.

Saez, Emmanuel, and Gabriel Zucman. 2019. *The Triumph of Injustice: How the Rich Dodge Taxes and How to Make Them Pay*. New York: W. W. Norton & Company.

Schmalbeck, Richard, and Jay A. Soled. 2009. "Elimination of the Deduction for Business Entertainment Expenses." *Tax Notes* 123 (6): 757–64

Sivadasan, Jagadeesh, and Joel Slemrod. 2008. "Tax Law Changes, Income-Shifting and Measured Wage Inequality: Evidence from India." *Journal of Public Economics* 92 (10–11): 2199–2224.

Smith, Matthew, Owen Zidar, and Eric Zwick. 2020. "Top Wealth in America: New Estimates and Implications for Taxing the Rich." Unpublished.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2019. "Capitalists in the Twenty-First Century." *Quarterly Journal of Economics* 134 (4): 1675–1745.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2020. "The Rise of Pass-Throughs and the Decline in the Labor Share." Unpublished.

Splinter, David. 2019. "U.S. Taxes are Progressive: Comment on Progressive Wealth Taxation." http://www.davidsplinter.com/Splinter-TaxesAreProgressive.pdf.

Splinter, David. Forthcoming. "U.S. Tax Progressivity and Redistribution." *National Tax Journal*.

Tazhitdinova, Alisa. 2020. "Are Changes of Organizational Form Costly? Income Shifting and Business Entry Responses to Taxes." *Journal of Public Economics*. 186: 1–13.

Thoresen, Thor O., and Annette Alstadsæter. 2010. "Shifts in Organizational Form under a Dual Income Tax System." *Finanz Archiv / Public Finance Analysis* 66 (4): 384–418.

Waseem, Mazhar. 2018. "Taxes, Informality and Income Shifting: Evidence from a Recent Pakistani Tax Reform." *Journal of Public Economics* 157: 41–77.

Wolfson, Michael, Michael Veall, Neil Brooks, and Brian B. Murphy. 2016. "Piercing the Veil: Private Corporations and the Income of the Affluent." *Canadian Tax Journal / Revue Fiscale Canadienne* 64 (1): 1–30.

Woodbury, Stephen A. 1983. "Substitution between Wage and Nonwage Benefits." *American Economic Review* 73 (1): 166–82.

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 52–78

Growing Income Inequality in the United States and Other Advanced Economies

Florian Hoffmann, David S. Lee, and Thomas Lemieux

The change in US income inequality over the last 40 years is one of the most extensively studied topic in economics. While it is well established that earnings and income inequality have increased sharply in the United States since the late 1970s, the explanations for the increase remain a matter of debate: for some examples in the literature, Goldin and Katz (2007) emphasize changes in returns to education; Acemoglu and Autor (2011) discuss the evolution of skills, tasks, and technologies; Acemoglu and Restrepo (2020) focus on robotization; and Fortin, Lemieux, and Lloyd (2019) consider the contribution of labor market institutions. No single explanation seems to be able to account for most of the growth in inequality. Indeed, the causes of rising inequality may differ across time periods and across middle, upper, and extreme upper income groups.

The purpose of this paper is three-fold. First, it documents key trends in US income inequality since the late 1970s, showing how much of the change comes from labor and non-labor market income. We will look at males and females separately, given the very different evolutions of their labor market participation during this time. In the case of non-labor market income, we focus on capital income in the form of interests, dividends, and (broadly defined) rental income. The empirical analysis is based on data from the March Annual Social and Economic Supplement

■ *Florian Hoffmann is Associate Professor of Economics, University of British Columbia, Vancouver, British Columbia, Canada. David S. Lee is Professor of Economics and Public Affairs, Princeton University, Princeton, New Jersey. Thomas Lemieux is Professor of Economics, University of British Columbia, Vancouver, British Columbia, Canada. Their email addresses are Florian.Hoffmann@ubc.ca, davidlee@princeton.edu, and Thomas.Lemieux@ubc.ca.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.4.52.

(ASEC) of the Current Population Survey (CPS) that has been collecting information about labor and capital income in a consistent fashion since 1976. To put these findings in context, pioneering work based on tax data by Piketty and Saez (2003) documents a dramatic increase in the concentration of income at the very top of the distribution. This initial evidence indicated that labor earnings were the main source of growth in top incomes. But with the labor share falling (Karabarbounis and Neiman 2014; Autor et al. 2020) and the continuing accumulation of wealth at the very top of the distribution (Saez and Zucman 2016), recent research has suggested that non-labor income has been playing an increasingly important role in inequality growth at the top (Piketty, Saez, and Zucman 2018). By contrast, evidence on the contribution of labor and non-labor income to the growth in income inequality among all income earners remains limited.[1]

Second, we assess the contribution of key explanatory factors, and in particular, education, to the growth in income inequality in the last four decades. While earlier research on income inequality using tax data provides excellent quality information on incomes at the top of the distribution, it contains limited information on the characteristics of tax filers. As a result, it offers little insight on how factors like education and occupation—which have been shown to play a major role in the labor literature—may also be affecting the distribution of non-labor income. Although education only accounts for a modest fraction of the *level* of earnings dispersion, it has been found to play a much larger role in the *growth* in earnings dispersion (Lemieux 2006a; Goldin and Katz 2007; Autor 2014). An important contribution of this paper is to study the connection between education and changes in the distribution of both labor and capital income. We also show that other explanatory factors—occupations, in particular—play a more limited role in inequality growth. This finding is consistent with the large literature showing that changes in the demand for different tasks, including but not limited to routine tasks, have contributed to the evolution of the wage distribution over time (for some entry points to this literature, see Acemoglu and Autor 2011; Autor and Dorn 2013; Firpo, Fortin, and Lemieux 2011; Caines, Hoffmann, and Kambourov 2017; Autor 2019).

Third, we compare the experience of rising US income inequality to other advanced economies. Existing studies show that earnings inequality has increased in Germany, the United Kingdom, and Italy but remained stable in France. We know little, however, about the role of capital income or about the relative contribution of education to inequality in these countries. Contrasting the evolution of inequality, and the source of the changes in inequality, in the United States and other high-income countries is helpful for understanding the factors behind these dramatic changes. In the early sections of this paper, we find that capital income has magnified the growth in US earnings inequality over time as the capital to labor income ratio disproportionately increased among high-earnings individuals. That

[1] Piketty, Saez, and Zucman (2018) look at the evolution of the share of income going to the bottom 50 percent and middle 40 percent of the distribution in addition to the top income shares, but there is a lot of dispersion within these broad groups that has not been as thoroughly studied.

said, labor income remains the main driver of inequality over the last 40 years, and it clearly would be difficult to slow down income inequality growth without addressing the inequality in labor income. We also find that education accounts for over half of the growth in US labor and capital income inequality. Growing income gaps among different education groups have led to a large expansion in between-group inequality, while the growing fraction of highly educated workers increased inequality because of composition effects. Other factors such as changing occupation premia and composition effects linked to the polarization of employment across occupations and space have also been playing a significant role in the growth in income inequality. Turning to large European economies, we show that inequality has been growing fast in Germany, Italy, and the United Kingdom, though not in France. As in the case of the United States, capital income only plays a limited role in inequality growth in these countries. Unlike the United States, income disparities linked to education is not a major factor in the rise in inequality in Europe, with the exception of Germany, where education can account for a substantial, though much smaller, part of the rise in income inequality.

Income Inequality Trends for the United States: Data and Measurement Issues

Our analysis of the trends in income dispersion in the United States is based on the IPUMS files of the March Supplement (ASEC) of the Current Population Survey (CPS) for 1976 to 2019, which collects income information for the preceding year (1975 to 2018).[2] The focus of this paper is on market income exclusive of taxes and transfers. The CPS contains information on net self-employment and wage and salary income over the reference year.[3] We define labor income as the sum of these two income sources. In the case of capital income, we combine income from three variables in the ASEC CPS: interest income; dividends; and rents, royalties, and income from estates or trusts.[4]

[2] For more detail on the IPUMS files, see Flood et al. (2000). Prior to the 1976 survey (income for 1975), the ASEC supplement only collected information at the individual level for heads of household. This is a major limitation because most female workers were not classified as household heads at the time. However, starting the analysis in the mid-1970s is not a significant limitation, given that inequality was relatively stable prior to about 1980.

[3] Smith et al. (2019) show that a large fraction of top incomes consists of entrepreneurial income earned through pass-through corporations (S-corporations and partnerships). Business owners may receive income in the form of wage and salary or business profit. In principle, both of these income sources should be captured in our CPS earnings measures that combine wages and salaries and net business income from the respondent's "own business" (what we refer to as self-employment income).

[4] Note that the ASEC CPS doesn't collect data on realized capital gains. This is a limitation, but for our present purposes not a major one. Many studies on the distribution of broader concepts of income often present results without capital gains (for example, Alvaredo et al. 2013) because of the high volatility of such gains over time, which in turn is linked to the fact investors may be strategic in deciding when to realize these gains.

Although most variables in the IPUMS files are fairly consistent over time, we made a few additional adjustments to the income variables, which are discussed in detail in the online Appendix, available with this article at the *JEP* website. Here, we briefly discuss two important adjustments.

First, for confidentiality reasons, the Census Bureau does not report incomes above a set threshold known as the top code. Between 1976 and 1995, incomes above the top code were simply replaced by the value of the top code (for example, $99,999 for wage and salary earnings in the late 1980s). Obviously, this made it difficult to use Census data to look at the top of the income distribution. The top coding procedure was improved in the 1996–2010 period by assigning a "replacement value" based on the average income of top-coded observations. After 2010, the Census Bureau moved to a "rank proximity swapping" procedure where high-income observations within a given range (above the top code) are swapped with close-by values and rounded off. Relative to earlier methods, the technique preserves the distribution above the top code and provides more accurate measures of the income distribution. Since then, the Census Bureau has provided swap values for years prior to 2011, which we use to keep income data consistent over time. As discussed in the online Appendix, the top earning shares for the top 1 and 10 percent that we calculate using the CPS are 9 and 34 percent, respectively, which is very similar to the shares found in tax data (see the updated version of the tables and figures from Piketty and Saez 2003, available at https://eml.berkeley.edu/~saez/TabFig2018.xls).

While this suggests that the Census Bureau's rank proximity swapping procedure approximates the upper tail of the earnings distribution reasonably well, it cannot fully adjust for changes in income data collection over time. As this inconsistency only affects the top 1 percent of earners, we trim that part of the distribution to make sure we have comparable measures of inequality over time in the analysis presented below. Note that using swap values remains important even when the top 1 percent is removed, as the fraction of observations with swapped values reaches up to 5 percent of the sample in some years.[5] We also remove observations with abnormally low average hourly earnings—less than $4 per hour in 2018 dollars—with the cut-off more or less corresponding to half of the real value of the minimum wage over the 1975–2018 period.

Second, income items other than earnings can be severely underreported in survey data (in this journal, Meyer, Mok, and Sullivan 2015). Rothbaum (2015) shows that only about 50 percent of capital income as measured in the national income

[5]The introduction of the computer-based questionnaire for the CPS in 1994 appears to have changed the upper tail of the distribution in a way that the Bureau of the Census rank proximity swapping procedure described above cannot fully account for The issue is discussed in detail in the online Appendix, available with this paper at the *JEP* website. Note that most women with swapped values are part of the (gender-specific) top 1 percent of earners. In the case of men, however, an average of 3 percent of earners—up to 5 percent in some years—have their earnings replaced with swapped values. As such, removing the top 1 percent of observations doesn't mitigate the importance of adjusting earnings using the Census Bureau's swapping procedure.

and product accounts gets reported in the CPS, in contrast with close to 100 percent of wage and salary earnings. Given the large underreporting of capital income, we adjust up reported capital income to match the figures from the national income and product accounts.[6] Note that although household members often share capital investments and their proceeds, the CPS collects information about capital income at the individual level, leaving it up to respondents to divide this source of income among themselves.

We focus our analysis on individuals from ages 25–64 who are working full-time/full-year in the reference year. The rationale for these sample restrictions is that we want to see how capital income contributes to overall income inequality for individuals with substantial labor income and who have been the focus of most of the earnings inequality literature. Many of the individuals under the age of 25 are still in school, and those who aren't haven't had much opportunity to accumulate savings. Likewise, most individuals over the age of 64 are retired, and only a modest share of their income comes from labor income. Given the substantially different trends in labor force participation, average earnings, and earnings dispersion for men and women, we follow the literature's typical practice of conducting the analysis in parallel for these two groups throughout the paper. Summary statistics for the sample, both broken by decade and pooled over all years from 1975 to 2018 using the CPS data, are available in the online Appendix.

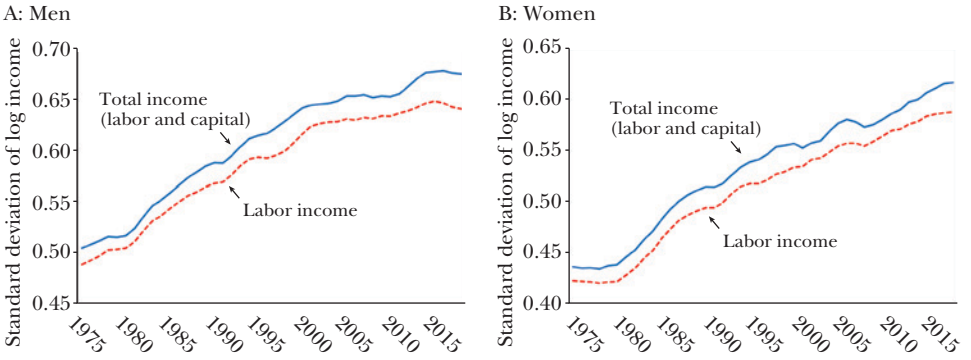
As our measure of inequality, we use the standard deviation of the log of income, which is a common metric in this literature. Some studies of inequality use the "coefficient of variation," which is the standard deviation divided by the mean. However, the distribution of income is of course a variable that is bounded by zero on the left and skewed to the right at the top levels. As a result, the mean will be well above the median. By using the log of income, our measure gives appropriately greater weight to lower and intermediate levels of income.

Inequality in the United States: Labor versus Capital Income

We take a first look at the contribution of both labor and capital income to overall inequality by contrasting the evolution of the standard deviation of log labor income and log total income in Figure 1. The gap between the two lines

[6]This adjustment is similar to the approach used by Piketty, Saez and Zucman (2018) to distribute some components of national income to households using a scaled-up version of survey self-reports. The adjustment factors we use are the inverse of underreporting ratios reported in Rothbaum (2015) for 2007–12: 1/.675 for interest income, 1/.695 for dividends, and 1/.274 for rents, royalties, and income from estates or trusts. While underreporting of capital income may be more severe in the upper part of the distribution, adjustments based on comparison of survey responses to aggregate figures—like the one used by Rothbaum (2015)—cannot be used to adjust for this potentially important issue. In the online Appendix, available with this paper at the *JEP* website, we also report inequality trends without this adjustment and discuss other changes in the survey instrument that may have improved the reporting of capital income in recent years.

Figure 1
Standard Deviation of Log Labor and Total Income



Source: Authors' calculations based on microdata from the March Annual Social and Economic Supplement (ASEC) of the Current Population Survey.
Note: The standard deviations are computed for a sample of full-time/full-year workers age 25-64 earning at least $4 per hour in 2018 dollars. The top 1 percent of the distribution has been trimmed because of inconsistencies in the way earnings at the very top have been collected over time. Labor income consists of net self-employment and wage and salary income. Total income is the sum of labor and capital income (interest income, dividends, and rents). See text for more detail.

represents the contribution of capital income. As mentioned above, these trends are computed for full-time/full-year workers, with the upper 1 percent of the distribution winsorized (that is, trimmed) to maintain data comparability over time. The figures are smoothed using a three-year moving average to facilitate the visual display. Although the three sources of capital income are combined together in this analysis, we note that most of the volatility in capital income is driven by interest and dividend income, with rental income remaining relatively stable over time. As there is no clear trend in the relative contributions of each source of capital income to total income inequality, we combine the three sources of capital income throughout the analysis.

In Figure 1, panel A shows that after growing modestly in the late 1970s, earnings inequality among men grew rapidly in the 1980s and 1990s. Interestingly, inequality then grew at a much slower pace after 2000. These trends are similar to those reported in earlier work (for example, Acemoglu and Autor 2011). Consistent with research based on tax data (for example, Piketty, Saez, and Zucman 2018), while capital income represents a modest fraction of total income, its distribution is substantially more skewed than the distribution of labor income.[7] Adding capital

[7]For our sample as a whole, men and women combined, close to 90 percent of capital income is concentrated among the top 10 percent, with 43 percent going to the top 1 percent. This is similar to findings from the tax data. For example, Saez and Zucman (2016) show that about 90 percent of wealth (or capital income) is held by the top 10 percent, and around 50 percent by the top 1 percent.

income to labor income leads to a higher level of dispersion for overall income (the blue line). Moreover, the contribution of capital income to the standard deviation of total income—the difference between the two curves in Figure 1—grows noticeably over time. The difference in standard deviations—with and without capital income included—grows from 0.014 in 1975–79 to 0.032 in 2014–18. The timing of changes in total income inequality is also substantially different from the ones for labor income only. The growth in total income inequality in the 1980s and 1990s is almost entirely driven by the increase in labor income inequality. By contrast, after year 2000, capital income plays an increasingly important role in overall inequality. These trends are qualitatively similar to studies of top incomes that have shown that while the increase in top shares was almost entirely driven by labor income in the 1980s and 1990s (as in Piketty and Saez 2003), capital income has been playing a more important role in recent years (as in Piketty, Saez, and Zucman 2018).

While trends for women shown in Figure 1b are generally similar to those for men, a few differences are worth noting. First, income inequality among women is completely flat in the late 1970s. A natural explanation for this difference relative to men is the minimum wage that was increasing during this period and had a larger impact on the inequality for women relative to men (DiNardo, Fortin, and Lemieux 1996; Lee 1999). Second, unlike in the case of men, earnings inequality among women keeps growing steadily after 2000. A possible explanation for this difference that we explore in the next section is that as the fraction of full-time/full-year women has been growing substantially over time, the composition of this group has also been changing in a way that resulted in more inequality.

Although we follow the literature in conducting the analysis separately for men and women, we note that inequality for men and women combined did not grow as fast as for men and women considered separately. To a great extent, this was driven by the decline over time in the between-group component of inequality linked to the gender gap.[8]

We further decompose the gap between labor income inequality and total income inequality into two components. The first is the *idiosyncratic* component, which represents the fact that individuals may have differing levels of capital income, even if they have the same labor income. The second is a *labor-correlated* component, which measures the extent to which total income inequality is magnified by the fact that higher-earning individuals tend to receive a larger share of total income from capital sources because higher-earning individuals would be expected to hold greater wealth, and thus, receive a higher share of their total income from capital sources.[9]

[8]This is illustrated in online Appendix Figure A1, available with this paper at the *JEP* website, which shows that the standard deviation of log total income grew by 0.12 between 1975 and 2018 for men and women combined, compared to 0.17 for men and 0.18 for women considered separately. However, we don't present results for men and women combined in the remainder of the paper, as doing so would complicate the decomposition exercises where we would have to keep track of how the different factors like education, occupations, parental roles, and others also affect the gender gap.

[9]Define total income Y as the sum of labor (Y_L) and capital income (Y_C). Log total income can be written as:

The evolution of these two components are shown in Figure 2. As in Figure 1, it is clear that the growing dispersion in labor income accounts for most of the increase in the variance of total income. At the same time, for both men (Figure 2, panel A) and women (Figure 2, panel B), the importance of the variance in the *labor-correlated* component of capital income increases steadily over time. This is directionally consistent with the finding of Piketty, Saez, and Zucman (2018) that capital income as a share of total income has disproportionately increased at the very top of the distribution. There is also some modest growth that is less pronounced in the *idiosyncratic* variation in capital income, indicating that capital income is getting more unevenly distributed conditional on labor income. Finally, the figure shows that labor income inequality is the main driver of the growth in the variance of total income until about 2000, at which point capital income becomes increasingly more important in accounting for total income inequality growth. For instance, after 2000, in the case of men, the *idiosyncratic* and *labor-correlated* components of capital income variance account for 15 and 24 percent of the growth in total income inequality, respectively.

In short, the growth in capital income inequality has been a nontrivial contributor to the growth in total income inequality over the last two decades, primarily driven by the fact that high-earnings individuals increasingly have a higher fraction of their incomes coming from capital income. This pattern of inequality change in the overall population of earners mirrors the findings of Piketty, Saez, and Zucman (2018) for the very top percentiles of earners, which indicates that the contribution of capital income in growing inequality extends beyond the very top of the distribution.

That said, the perspective provided by Figures 1 and 2 makes clear that the long-run growth in total income inequality over the past several decades is driven primarily by growth in the labor income inequality. In fact, using the trends in labor income inequality to proxy for the magnitude of the growth in total income inequality does a reasonable job, whereas the same could not be said about the trends in capital income inequality. With this as context, we take advantage of the rich set of individual characteristics available in the CPS data to look at the contribution of various factors, and education in particular, in the growth of total income inequality.

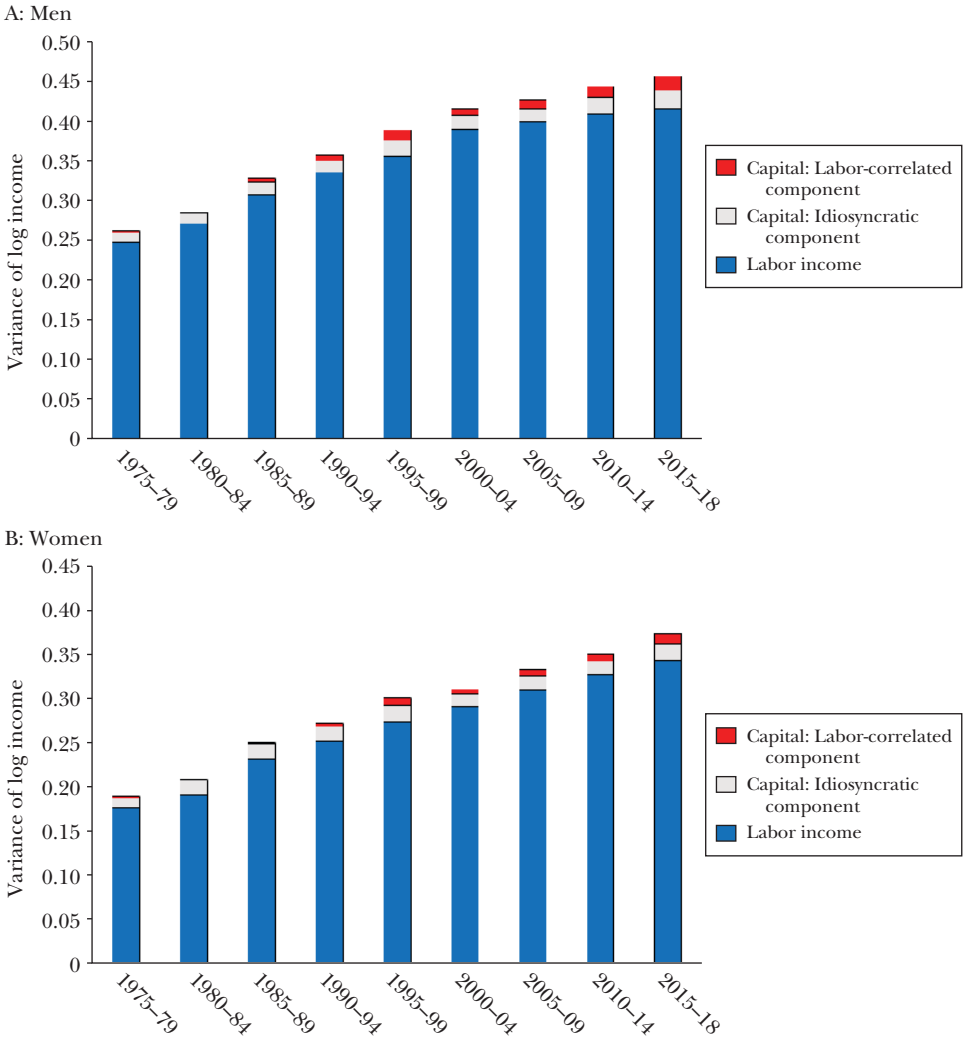
$$\log(Y) = \log(Y_L + Y_C) = \log(Y_L) + \log(1 + r) \approx \log(Y_L) + r,$$

where $r = Y_C/Y_L$ is the ratio of capital to labor income, and we use the fact that $\log(1 + r) \approx r$ for small values of r. To simplify the exposition, we replace $\log(1 + r)$ with r hereinafter, but we do keep using $\log(1 + r)$ for computations. We also use small caps to denote the log incomes $y = \log(Y)$ and $y_L = \log(Y_L)$. The contribution of these two factors to total income dispersion can be formally obtained using a variance decomposition:

$$\text{Var}(y) = \text{Var}(y_L) + \text{Var}(r) + 2\,\text{Cov}(y_L, r).$$

$\text{Var}(r)$ is the idiosyncratic component that captures variation in capital income that is unrelated to labor income; $2\,\text{Cov}(y_L, r)$ is the labor-correlated component that captures the systematic relationship between r and labor income.

Figure 2
Variance Components of Total Income (Labor and Capital)



Source: Authors' calculations based on microdata from the March Annual Social and Economic Supplement (ASEC) of the Current Population Survey.
Note: See the note to Figure 1 for details on the sample. The *idiosyncratic* component of capital income reflects the variation in capital income among individuals with the same labor income. The corresponding variance component is computed as the variance of the ratio of capital to labor income. The *labor-correlated* component of capital income captures the extent to which total income inequality is magnified by the fact that higher-earning individuals tend to receive a larger share of total income from capital sources. The corresponding variance component is computed as (twice) the covariance between log labor income and the ratio of capital to labor income. See text for more detail.

The Role of Education in Inequality Growth

Rates of returns to education have increased substantially since the late 1970s. In their highly influential study, Katz and Murphy (1992) link the sharp growth in the college wage premium during the 1980s to a deceleration of the growth in the relative supply of college education in an era where the relative demand for highly educated workers was increasing. Numerous other studies have shown that the returns to education kept increasing after the 1980s (for example, Card and Lemieux 2001; Goldin and Katz 2008; Acemoglu and Autor 2011; Autor 2014). Fewer studies have sought to quantify the contribution of education to the overall growth in income inequality, but those studies suggest that it may have played a disproportionately large role in the growth in dispersion of earnings. For instance, Lemieux (2006a) and Goldin and Katz (2007) find that at least one-half of the growth in earnings dispersion can be connected to growing returns to education. The twin goals of this section are to evaluate whether this finding still holds when using more recent data and whether education plays an important role in the dispersion of both labor and capital income.

The CPS data give us the ability to look at how income varies with respect to "groups" defined by education, age, or labor market experience, as well as occupation or industry, and how income for those groups evolves over time. For any defined grouping, we can decompose overall income dispersion into three components: (1) *between-group* inequality, (2) *within-group* inequality, and (3) *composition* effects. The literature uses these concepts to decompose the change in inequality over time. An example of rising *between-group* inequality is when the income gap between high- and low-educated workers widens, which naturally will lead to increases in the gap between high- and low-income workers overall. Rising labor market returns to education will increase the between-group component of inequality when groups are defined by education.

Rising *within-group inequality* occurs when the gap between high- and low-income workers widens even for people in the same "group." For example, there is a fair amount of variability in income among workers who have a college degree, potentially driven by varying quality of the college education itself. So a growing demand for workers from colleges of higher quality could be driven by increases in *within-group* inequality.[10] Another possible source for growing within-group dispersion among college-educated workers is that the demand for their skills may be growing unevenly across space. For instance, Autor (2019) shows that the college wage premium has grown much faster in high- relative to low-density urban areas. Autor also shows that this phenomenon is connected to a faster growth in the

[10] See Lemieux (2006a) for a formal exposition of this argument in a context where returns to education are heterogenous across individuals. If school quality is the source of heterogenous returns, an increase in the demand for effective education skills (both quantity and quality) will lead to an increase in both the college–high school gap and in earnings dispersion among college-educated workers.

demand for high-skill tasks (professional, technical, and managerial occupations) in high-density urban areas.

Finally, *composition* effects arise simply because, if over time, there are more and more people in groups that tend to have more *within-group* dispersion, this by itself will increase overall income inequality. For example, over the past several decades, there has been a steady shift in the proportion of the workforce from lower to higher education levels, which then would tend to lead to more inequality because income is more dispersed among highly educated workers (Lemieux 2006b).

We capture these three potential contributions of education to the growth in the variance of log income using a variance decomposition. We illustrate the results in Figure 3, again dividing into men and women workers. Groups used for the decomposition are formed using five education groups and eight age groups, which in this literature are often used as a proxy for the level of job experience. Note that in the case of between-group inequality, we compute the contribution of both education- and experience-related wage differentials to inequality growth. Within-group inequality is broken down into a base component capturing income dispersion among individuals with less than a college degree—the "high school" group—and the difference in within-group inequality between those with or without a college degree. The latter represents the contribution of education to the growth in within-group inequality.[11]

Figure 3 shows that within-group dispersion (represented by the bars labeled "within for HS" and "effect of education on within") accounts for most of overall income dispersion during each time period. The "within for HS" bar represents the within-group variance for the high-school group, while the "effect of education on within" bar reflects that the within-group variance for college-educated workers is

[11] For a sense of our approach, consider a Mincer-type log income regression where C is a dummy for college, δ is the return to college, and u is an error term capturing unmeasured ability and, as discussed above, college quality or differences in returns to college across space:

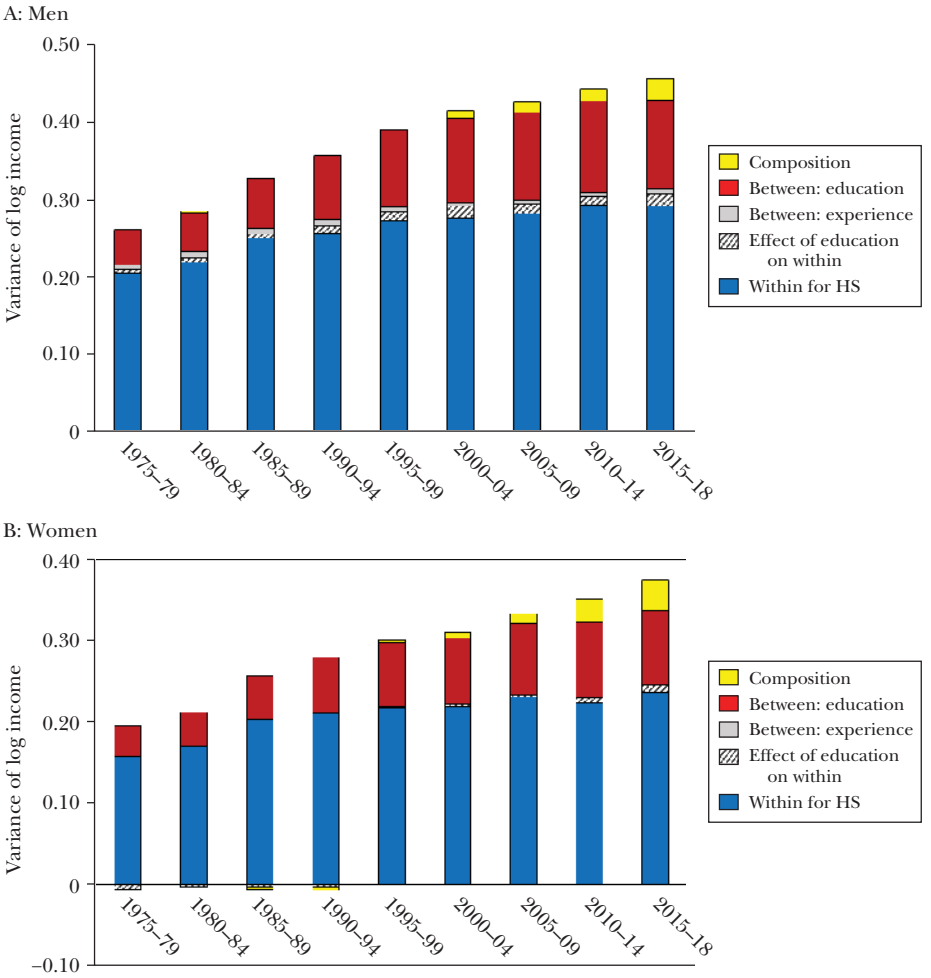
$$y = \delta C + u.$$

Let θ represent the fraction of individuals with a college degree, σ_C^2 represents the variance of u among college-educated individuals, and σ_H^2 represents the variance of u among high school-educated individuals. The variance of income can be written as the sum of the between-group component, $\delta^2\theta(1-\theta)$, and within-group component, $\theta\sigma_C^2 + (1-\theta)\sigma_H^2$. Adding a time subscript and re-arranging terms yields:

$$\mathrm{Var}_t(y) = \delta_t^2\,\theta_t(1-\theta_t) + \theta_t(\sigma_{Ct}^2 - \sigma_{Ht}^2) + \sigma_{Ht}^2.$$

When looking at changes in the variance from a base period 0 to time t, the contribution of the three factors discussed in the text can be obtained as follows. To compute the contribution of composition effects, we replace the college share θ_t by its value in the base period, θ_0, which amounts to re-weighting college workers using the reweighting factor θ_0/θ_t. For between-group effects, we recompute the equation by replacing δ_t^2 with δ_0^2. For within group effects (excess growth for college relative to high school-educated individuals), we replace $\sigma_{Ct}^2 - \sigma_{Ht}^2$ with $\sigma_{C0}^2 - \sigma_{H0}^2$. After having done these adjustments, the only source of growth left in the variance is the change over time in the within-group variance of high school-educated individuals $(\sigma_{Ht}^2 - \sigma_{H0}^2)$, which represents the baseline change if education played no role in the change in inequality. While in the text we only use two education groups to simplify the exposition, in the empirical analysis we use five education groups (high school dropouts, high school graduates, some college, college graduates, and college post-graduates). As a proxy for shifts in experience levels, we also control for age, using dummies for five-year age categories going from 25–29 to 60–64.

Figure 3
Sources of Change in the Variance of Log Total Income



Source: Authors' calculations based on microdata from the March Annual Social and Economic Supplement (ASEC) of the Current Population Survey.
Note: See the note to Figure 1 for details on the sample. The within for HS component is the variance of log total income among high-school graduates (individual with less than a college degree). The effect of education on within component is the difference between the variance of log income for college and high-school graduates. The between: experience component is the between-group variance linked to experience-related wage differentials. The between: education component is the between-group variance linked to education-related wage differentials. The composition component represents the change in the variance of log total income linked to changes in the distribution of education and experience relative to the base period (1975–79).

larger; this latter component, especially in the earlier time periods like 1975–79, is relatively small throughout the entire period. As is well known—for example, from Juhn, Murphy, and Pierce (1993)—within-group dispersion grew substantially

during the 1980s, accounting for a substantial share of the growth in the variance of income. However, most of the growth in within-group dispersion stopped after the 1985–89 time period.

By contrast, between-group dispersion (the red and grey bars) grew over the entire 1975–2018 period. Almost all of the between-group dispersion at a given point in time is linked to education (the red bar) rather than experience (the gray bar). The size of this education-related variance component more than doubled: for men, from 0.045 in 1975–79 to 0.115 in 2015–18; for women, from 0.038 in 1975–79 to 0.091 in 2015–18. Figure 3 also indicates that, after 1985–89, this variance component played a larger role in the growth in income dispersion than did the within-group component unrelated to education (baseline within-group dispersion for high school-educated workers).

Figure 3 reveals that most of the inequality growth after 1985–89 is due to the sum of three variance components linked to education: 1) the *between-education-group* dispersion (the red bar), 2) the growth in *within-group* inequality for college-educated workers *over and beyond* the growth in the within-group inequality for high school-educated workers, and 3) finally, particularly starting in 2000, *composition* effects linked to the shift from high school-educated to college-educated workers.

We quantify the role of education in inequality growth by showing the contribution of each variance component in percentage terms in Table 1. For the entire 1975–79 to 2015–18 period, we show the decomposition both for total income and labor income only. The results are very similar, indicating that education makes a similar contribution to the growth in total income or labor income only (a figure showing the labor income decomposition by five-year intervals is also available in the online Appendix, available with this article at the *JEP* website). Table 1 confirms that most of the growth in income inequality over the 1975–79 to 2015–18 period—56 percent for both men and women—is connected to education. The fraction grows even higher—around 70 percent—when only focusing on changes that occur after the late 1980s. Most of the contribution of education is due to the between-group component and composition effects, with the effect of education on within-group dispersion playing a more minor role. Composition effects play a more significant role for women, while the growth in the between-group component is larger for men. The former is not surprising since the composition of the female workforce has dramatically changed over time, with the fraction of women with a college degree increasing from 0.192 in 1975–79 to 0.469 in 2015–18. The growth in the educational attainment of men has been more moderate, with the fraction of college-educated workers going up from 0.246 in 1975–79 to 0.393 in 2015–18. Because within-group dispersion is now higher among college than high school graduates, the faster growth in college-educated labor among women leads to larger composition effects. We also note that while the composition effects reported here combine the contribution of education and experience, 80 percent of composition effects for men and 87 percent of composition effects for women over the 1975–1979 to 2015–2018 period are due to education only.

Table 1

Contribution (in %) of Education and Other Factors to the Growth in the Variance of Total Income

		Contribution of education			
	Within (HS)	*Between: experience*	*Between: education*	*Education effect on within*	*Composition effects*
A. Men					
Total income:					
1975–79 to 1985–89:	67.8	2.3	28.8	0.9	0.2
1985–89 to 2015–18:	32.2	–2.0	39.6	8.9	21.4
Total change:	44.4	–0.6	35.9	6.2	14.1
Labor income:					
Total change:	45.4	–0.9	37.5	6.3	11.7
B. Women					
Total income:					
1975–79 to 1985–89:	74.6	0.6	26.2	4.3	–5.7
1985–89 to 2015–18:	27.2	0.6	30.1	9.5	32.6
Total change:	42.8	0.6	28.8	7.8	20.0
Labor income:					
Total change:	42.1	0.1	30.5	8.3	18.9

Source: Authors' calculations based on microdata from the March Annual Social and Economic Supplement (ASEC) of the Current Population Survey.
Note: See the note to Figure 1 for details on the sample and the note to Figure 3 for an explanation of the variance components presented in the table.

In summary, most of the growth in labor and capital income inequality can be linked to education. In particular, increasing returns to education lead to a large increase in the between-group component that accounts for around one-third of the increase in the variance of income between 1975–79 and 2015–18. Another important factor linked to education is that since incomes are more unequally distributed among college-educated workers, the growth in the fraction of highly educated workers leads to large composition effects, especially among women. The faster growth in within-group dispersion among college-educated workers also contributed, albeit in a more modest way, to the increase in overall income inequality. If it had not been for factors directly connected to education and increasing gains to education, income inequality would have increased by less than half as much as it did over the last four decades.

The Role of Occupation, Industry, and Location

In this section, we compare the role of education documented above to that of occupation, industry, and location in accounting for the level and growth of

total income inequality. There is a rich literature looking at how relative changes in the demand for labor by industry and occupation have been important factors in growing returns to education, and to inequality more generally. A group of papers in the early 1990s sought to explain, using skill-biased technical change or related concepts, the monotonic relationship between skill level and earnings changes that was observed during the 1980s. For instance, Bound and Johnson (1992) and Katz and Murphy (1992) use "shift-share" approaches to look at whether the relative growth in industries employing more educated labor has contributed to the growth in the rate of return to education. Other papers such as Krueger (1993) and Berman, Bound, and Griliches (1994) argue that the growing returns to education were primarily due to skill-biased technical change linked to the computer revolution.

Starting in the 1990s, however, inequality growth became increasingly concentrated at the top of the distribution. Furthermore, earnings at the bottom end of the distribution stabilized relative to those earnings in the middle, leading to what Autor, Katz, and Kearney (2006) famously called the polarization of the earnings distribution. This phenomenon is also present in the CPS data used in this paper, which shows that for both men and women, the gap between the 90th and the 50th percentiles grew steadily since the late 1970s. By contrast, all of the growth in the gap between the 50th and the 10th percentiles is concentrated in the 1980s.[12]

Autor, Katz, and Kearney (2006, building on Autor, Levy, and Murnane 2003) conjectured that a more nuanced form of technological change could explain the polarization of earnings of the 1990s: specifically, computerization might have a particularly negative impact on routine tasks that used to be performed by workers in the middle of the income distribution. This insight changed the focus of the inequality literature from industries to occupations, as occupations are much better proxies for the types of tasks performed by workers of different skill levels. Numerous studies have shown that, consistent with the "routine-biased" technical change hypothesis, the distribution of employment across occupations has become increasingly polarized in the United States and other advanced economies (for example, Goos, Manning, and Salomons 2014).

More recently, Autor (2019) introduced an important new dimension to employment polarization by showing that the distribution of occupations performed by workers of different skill levels has changed substantially across place during the last few decades. Autor shows that non-college workers used to disproportionately hold middle-skill jobs—blue-collar production and white-collar office jobs—in densely populated urban areas. These non-college urban workers were hit particularly hard by routine-replacing technical change. Autor shows that this changing distribution of employment over both occupation and space played an important role in inequality growth.

[12] For an illustration of this pattern, see Appendix Figure A3 available with this paper at the *JEP* website.

Trade and globalization may also have contributed to the polarization of the labor market. For instance, Chetverikov, Larsen, and Palmer (2016) find that low-wage earners were significantly more affected by increased Chinese import competition—what Autor, Dorn, and Hanson (2013) called the "China shock"—than high-wage earners.

We measure these developments within our variance decomposition framework by first adding variables for occupation, industry, and location. Our objective here is to assess how much of the rise in income dispersion can be explained by these factors, above and beyond what is already being explained by education. This is accomplished by further refining the groups—which were in Figure 3 limited to education and experience—to reflect additionally occupation, industry, and location. We note that this calculation may understate the full contribution of changing demand by occupation, industry, and location, because it does not capture the part of the contribution that is being mediated through education. We also explore how adding these factors changes the magnitude of the composition effects shown in Figure 3. Like Autor (2019), we use DiNardo, Fortin, and Lemieux's (1996) reweighting method to compute a shift in the composition of the labor market compared with the counterfactual income distribution that would have prevailed if the distribution of occupation and place had remained unchanged since the late 1970s.

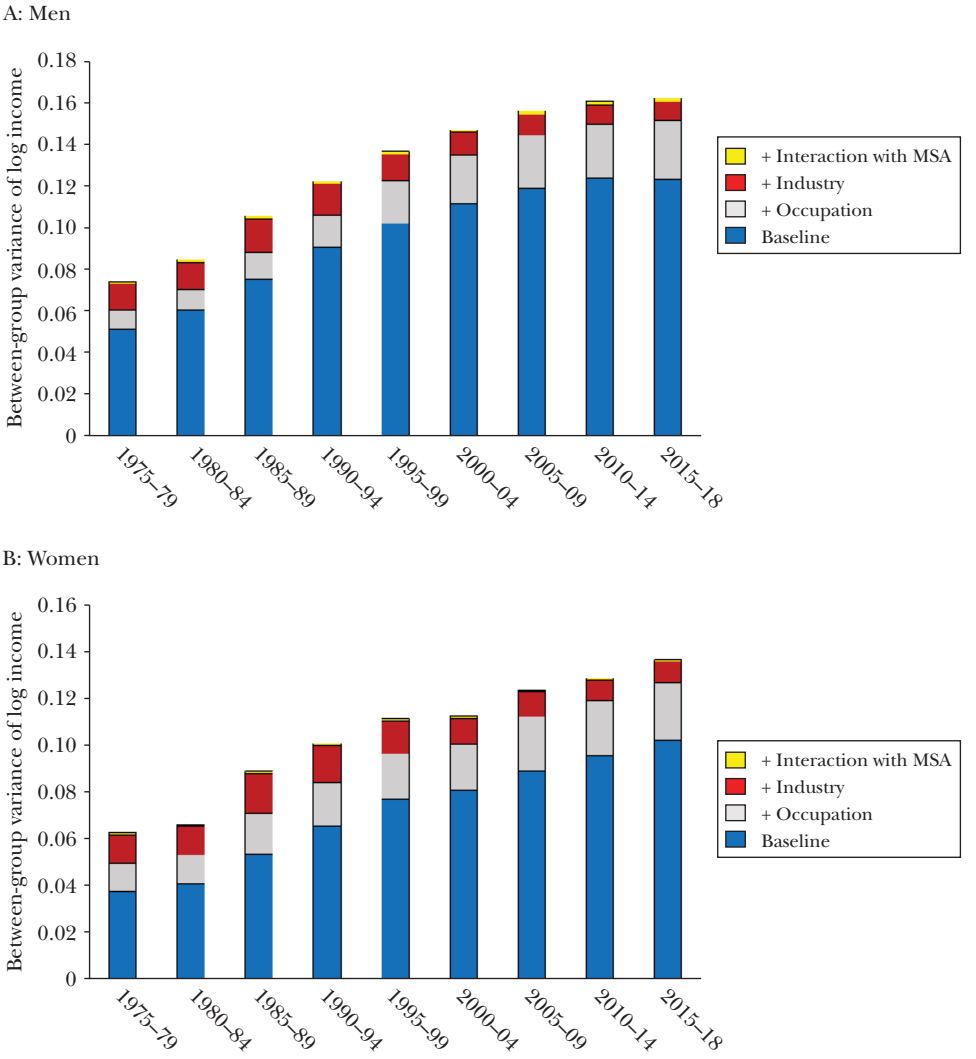
Occupations are coded up using the same nine categories as Autor (2019). In the case of industries, we classify workers into 12 broad categories based on the 1990 Standard Industrial Classification harmonized over time. Regarding the spatial distribution of workers, we use a classification based on whether individuals live in (1) the 15 most populous metropolitan statistical areas, (2) other metropolitan statistical areas, or (3) non-urban areas.[13]

Figure 4 shows the effect of adding more covariates on the between-group variance of total income. The focus on the between-group component explains why the variances reported in Figure 4 are substantially lower than those reported in the previous figures. The baseline (lower blue bar) reproduces the sum of the two between-group variance components based on education and age in Figure 3. For both men and women, adding occupation, industry, and location appears to explain substantially more of the variance in total income at any given point in time. For example, in the case of men, adding these factors raises the total between-group variance component from about 0.05 to about 0.075 in 1975–79, and from 0.12 to 0.16 in 2015–18.

As for how these factors contribute to the overall *growth* in total income inequality, Figure 4 shows two patterns. First, the between-group variance component linked to industry (the red bar) has been declining over time, and changes linked to earnings changes over space are small (as also documented in Autor 2019), compared to the between-group changes linked to education, for example.

[13] Again, for more details on these variables and full methodology behind the calculations, see the online Appendix, available with this paper at the *JEP* website.

Figure 4
Effect of Additional Covariates on the Between-Group Variance



Source: Authors' calculations based on microdata from the March Annual Social and Economic Supplement (ASEC) of the Current Population Survey.
Note: See the note to Figure 1 for details on the sample. The *baseline* component represents the between-group variance of log total income due to experience- and education-related wage differentials only. The *occupation* component indicates by how much the between-group variance increases when occupation-related wage differentials are taken into account by adding occupation dummies to a log income regression with a full set of education times experience effects. Likewise, the *industry* component indicates by how much the between-group variance increases when industry dummies are further added to the log income regression. Finally, the *interaction with MSA* component indicates by how much the between-group variance increases when a set of MSA dummies and its interaction with occupation, industry, and education dummies are added to the log income regression.

Second, Figure 4 confirms existing findings that occupation wage differentials have been playing an increasingly important role in income inequality growth. Acemoglu and Autor (2011) reach a similar conclusion by including occupation dummies in a Mincer-type equation. Likewise, Firpo, Fortin, and Lemieux (2011) and Fortin and Lemieux (2016) show that either occupation dummies, or characteristics of occupations summarized by tasks measures, contribute to the growth in the between-group variance.

Table 2 quantifies the extent to which the additional consideration of occupation and location can account for the growth in total income inequality. For the sake of brevity, we only show changes over the whole 1975–79 to 2015–18 period. The first row in each panel (A–D) uses only education and age to define the groups, while the second row additionally includes occupation and metropolitan statistical area so that the difference between the two quantifies the importance of the occupational and locational dimension. The first column reports the overall change in inequality, matching the numbers in Figure 2. The second column reports the between-group components of variance as illustrated in Figure 4. It shows that for both men and women, and for the total income (panels A and B) and labor income only measures (panels C and D), occupation and location contribute an extra 0.015 to 0.020 relative to a base of 0.059 to 0.072 explained by education and age alone. The decomposition in Figure 4, with its focus on between-group variance components, did not allow for composition effects. So the third and fourth columns use a re-weighting approach (as was used to produce Figure 3 to compute the composition effects components).

Interestingly, the contribution of the between-group component declines when we add occupation and metropolitan statistical area but is offset to varying degrees by the composition effects. This finding reflects a subtle interaction between the composition of the workforce and the magnitude of the effect of different factors on income. To the extent that returns to high levels of education and high-paying occupations have grown over time, downweighting the importance of these groups by holding the occupational distribution fixed at the 1975–79 level dampens the contribution to the between-group component to the growth in income inequality.

The additional consideration of occupation and metropolitan statistical area seems to make the most difference via composition effects in the case of women, for whom the contribution of composition effects to growth in the variance of total income increases from 0.037 to 0.065. The latter figure represents more than one-third of the growth in the overall variance between 1975–79 and 2015–18. This is consistent with Autor's (2019) finding that the spatial and occupational polarization of work has played an important role in the secular increase in income inequality.

Although occupation and place play an interesting role in the evolution of income inequality over time for women, in the case of men, they add only modestly to what can already be explained using only education (and experience). The final column of Table 2 reports the ratio of the sum of the components in the third and fourth columns to the first column. It shows that adding occupation and space does

Table 2

Change in the Variance of Total Income between 1975–79 and 2015–18: Contribution of Between-Group and Composition Effects with Different Set of Covariates

		No reweighting	*With reweighting*		
	Total change	*Between-group*	*Between-group*	*Composition*	*% explained*
A. Men, total income					
Education*Age	0.1949	0.0720	0.0689	0.0275	49.5
+Occupation*MSA	0.1949	0.0882	0.0627	0.0358	50.5
B. Women, total income					
Education*Age	0.1850	0.0647	0.0543	0.0371	49.4
+Occupation*MSA	0.1850	0.0750	0.0414	0.0650	57.5
C. Men, labor income					
Education*Age	0.1676	0.0616	0.0614	0.0196	48.3
+Occupation*MSA	0.1676	0.0764	0.0549	0.0274	49.1
D. Women, labor income					
Education*Age	0.1659	0.0588	0.0508	0.0314	49.5
+Occupation*MSA	0.1659	0.0681	0.0379	0.0582	57.9

Source: Authors' calculations based on microdata from the March Annual Social and Economic Supplement (ASEC) of the Current Population Survey.
Note: See the note to Figure 1 for details on the sample. The rows labelled *Education*Experience* indicate the change in the between-group variance due to education- and experience-related wage differentials only. The rows labelled + *Occupation*MSA* also consider the contribution of occupations interacted with MSA effects. The column *Between-Group with reweighting* shows the change in the between-group variance when the distribution of the explanatory variables (education and experience with and without occupation and MSA effects) is held constant at its 1975-79 level. The *Composition* column shows how much of the increase in the variance is due to changes in the explanatory variables.

not drastically change how much of the growth in the variance of income (about 50 percent) can be accounted for by the between-group component and composition effects. By contrast, the additional explanatory effect of considering occupations plays a more important role for women, raising the percentage from about 50 to 58 percent. This likely reflects the fact that the distribution of occupations has been changing more drastically for women than men over time, with women increasingly moving into high-paying managerial and professional occupations that were dominated by men back in the late 1970s.

Evidence for Large European Economies

Many of the explanations for the growth in income inequality in the United States, such as those based on technological change and employment polarization, should also apply to other high-income economies. Back in the 1990s, a major challenge to this view was that inequality had only grown modestly, if at all, in most other

advanced economies. For instance, Freeman and Katz (1995) show that, unlike in the United States, inequality was relatively stable in most European economies and Japan during the 1980s. The only notable exception was the United Kingdom where, like in the United States, inequality grew rapidly during the 1980s; indeed, Machin (2011) shows that inequality continued to increase steadily over time in the United Kingdom, albeit at a faster rate during the 1980s. Freeman and Katz (1995) suggest that a combination of differences in national wage-setting institutions and supply factors (especially the rate of growth in highly educated labor) could go a long way towards explaining these differences.

However, more recent studies indicate that earnings inequality has been increasing in several continental European countries since at least 1990. For example, Dustmann, Ludsteck, and Schönberg (2009) use high-quality social security data to show that earnings inequality has been steadily growing in Germany over the last few decades, thereby revising the findings of a stable earnings distribution from earlier studies that were based on the German Socio-Economic Panel (GSOEP) (for example, see Steiner and Wagner 1998). Card, Heining, and Kline (2013) and Hoffmann (2019) offer a further analysis of the growth in inequality in Germany. For Italy, Manacorda (2004) shows that inequality started growing in the late 1980s after a wage indexation mechanism known as the *Scala Mobile* became much less binding. Devicienti, Fanfani, and Maida (2019) show that inequality kept growing steadily in Italy after the end of the analysis period considered by Manacorda (from 1977 to 1993).

We reexamine these trends using the most recently available data for France, Germany, Italy, and the United Kingdom. As in the case of the United States, our focus is on documenting trends for both labor and total income. This presents an empirical challenge because none of these four large European countries collect annual data providing the detailed information about income and individual characteristics that is contained in the US March CPS.[14] In an effort to maximize comparability with the US results, we rely on the Household Budget Survey for France, the Survey of Household Income and Wealth (SHIW) for Italy, the Family Resources Survey (FRS) for the United Kingdom, and the Income and Expenditure Survey (EVS: Einkommens- und Verbrauchsstichprobe) for Germany.

In the case of French, Italian, and UK data, we use the harmonized version of these data provided by the Luxembourg Income Study (LIS 2020) project. However, since the Luxembourg Income Study relies on the GSOEP for Germany, which seems to miss (as discussed above) some of the inequality trends found in high-quality administrative data, we use the EVS data (the Sample Survey of Income and Consumption) provided by the German Statistical Office instead. The EVS data are

[14] Numerous studies of European economies have used rich longitudinal social security data sets to look at labor market inequality, including Dustmann, Ludsteck, and Schönberg (2009); Card, Heining, and Kline (2013); and Hoffmann (2019) for Germany, and Devicienti, Fanfani, and Maida (2019) and Daruich, Di Addario, and Saggio (2020) for Italy. However, these data sets don't provide information on non-labor income and on workers who are not covered by social security (self-employed and public sector workers in Germany).

collected administratively, and among other uses, they determine the consumption basket for the calculation of the official consumer price index and for calculating the income thresholds of unemployment and social insurance. Germany's Federal Statistical Office explicitly highlights its high accuracy. Indeed, Dustmann, Fitzenberger, and Zimmermann (2018), in a study of the evolution of inequality at the household level, find that inequality trends in the EVS track closely those documented in administrative social security data.[15] Two major advantages of the EVS relative to administrative social security data are that they report capital income and that their top-code is high.

We provide more information on these data sets and discuss their limitations relative to the US CPS in the online Appendix, available with this article at the *JEP* website. Two important differences of the European data sets worth mentioning here are: (1) capital income is only collected at the household level, not the individual level; and (2) in the European data sets, full-time status is more frequently available than information about weeks of work and full-year status. We adjust for the first issue by dividing household capital income by the number of individuals age 25–64 in the household. We address the second issue by only keeping years where full-time status is available.[16] Also, to mimic our sample restriction in the US data of keeping only full-time/full-year workers earning more than $4 per hour, we remove all observations with annual earnings below $8,000 ($4 times 2,000 hours a year) in 2018 terms from our European data.

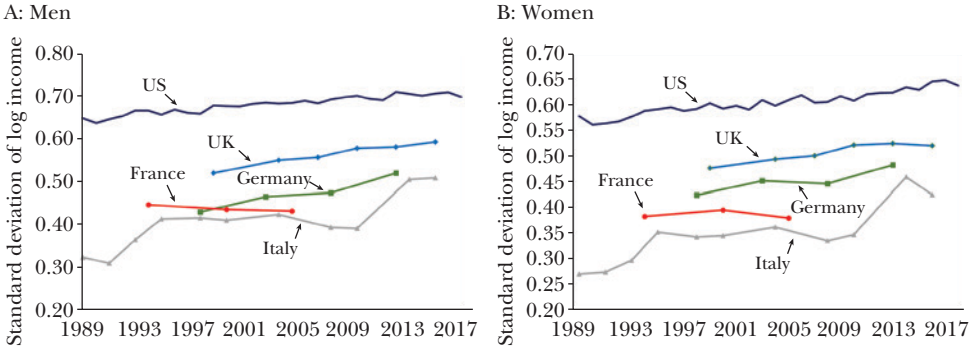
Figure 5 shows the evolution of the standard deviation of log total income in European countries and in the United States. We show the trends starting in 1989, the first year for which European data are available. For the sake of comparability, we use the full-time/over $8,000 sample criterion in US data, too, instead of the full-time/full-year criterion used in prior tables and figures. Comparing Figures 1 and 5 indicates that the US standard deviation grows somewhat more slowly when using the full-time/over $8,000 criterion instead of full-time/full-year, though the overall trends remain similar. For example, in the case of men, the standard deviation increases by 0.083 between 1989 and 2018 in Figure 1, panel A, compared to 0.050 in Figure 5, panel A.

Figure 5 shows that, for both men (Figure 5, panel A) and women (Figure 5, panel B), income inequality has increased in all countries but France since the 1990s. While we are unable to analyze data from France after 2005, other studies using slightly different samples and income concepts have generally found that inequality has remained fairly stable in France since 2005; for example, Boiron (2016) uses the French Household Budget Survey data to study the evolution of income inequality without imposing the full-time/over $8,000 restrictions and has

[15] More details on the strengths and weaknesses of the different German data sets is provided in the online Appendix, available with this paper at the *JEP* website. Also, we discuss in more detail some new insights on the evolution of inequality in Germany in Appendix B.

[16] Unfortunately, we have to drop the most recent observation (2010) for France and pre-1998 observations for Germany because of the lack of information about full-time status in those years.

Figure 5
Standard Deviation of Log Total Income in European Countries



Source: Authors' calculations based on microdata from the March CPS for the United States, the Household Budget Survey for France, the Survey of Household Income and Wealth for Italy, the Family Resources Survey for the United Kingdom, and the Income and Expenditure Survey for Germany.
Note: The standard deviations are computed for a sample of full-time workers age 25-64 with annual earnings of at least $8000 in 2018 dollars (adjusted for exchanges rates in the case of European countries). Total income is the sum of labor (net self-employment and wage and salary income) and capital income (interest income, dividends, and rents). The top 1 percent of the distribution (top 1.5% in Germany for reasons explained in the online Appendix) has been trimmed because of inconsistencies in the way earnings at the very top have been collected over time in the United States. See text for more detail.

access to a wider time period than what is available in the Luxembourg Income Study data. He finds that both the Gini coefficient and the 90/10 ratio as measures of income inequality have been essentially unchanged in France between 2005 and 2013. Thus, France appears to be increasingly an outlier relative to the three other large European economies where inequality has increased rapidly over the last few decades. And although the level of inequality remains higher in the United States, the inequality gap relative to Germany, Italy, and the United Kingdom has declined in recent decades as inequality has grown slightly faster in these three countries than in the United States.

Unlike in the United States, the evolution of total income inequality in Europe is almost entirely driven by changes in the distribution of labor income, and capital income plays a very small role. Indeed, the difference between the variance of total and labor income is an order of magnitude smaller in all European countries than in the United States.[17] With respect to the role of capital income in the

[17] For calculations, see online Appendix Table A2, available with this paper at the *JEP* website. We don't separately show the evolution of labor and total income in Figure 5 to avoid overloading the graphs, but online Appendix Table A3 shows that even when no adjustment is used to reconcile capital income as reported in the March CPS with measures of capital income from the national income and product accounts, capital income still plays a more important role in the United States than in European countries.

Table 3
Contribution (in %) of Education and Other Factors to the Growth in the Variance of Total Income

			Contribution of education			
	Within (HS)	*Between: experience*	*Between: education*	*Ed. effect on within*	*Composition effects*	*Total change (by decade)*
A. Men						
US: 1985–89 to 2015–18	32.2	–2.0	39.6	8.9	21.4	0.043
France: 1994 to 2005	13.7	–26.0	171.6	40.0	–99.2	–0.011
Italy: 1989 to 2016	54.3	5.1	7.0	4.8	28.7	0.057
Germany: 1998 to 2013	49.8	–0.9	15.7	11.2	24.2	0.060
UK: 1999 to 2016	47.3	–2.5	–12.6	18.7	49.1	0.048
B. Women						
US: 1985–89 to 2015–18	27.2	0.6	30.1	9.5	32.6	0.041
France: 1994 to 2005	373.0	–101.3	324.6	163.3	–659.6	–0.002
Italy: 1989 to 2016	46.0	3.8	9.6	12.3	28.2	0.040
Germany: 1998 to 2013	22.8	5.7	19.2	24.2	28.1	0.039
UK: 1999 to 2016	49.3	–5.8	–51.2	35.1	72.5	0.025

Source: Authors' calculations based on microdata from the March CPS for the United States, the Household Budget Survey for France, the Survey of Household Income and Wealth for Italy, the Family Resources Survey for the United Kingdom, and the Income and Expenditure Survey for Germany.
Note: See the note to Figure 5 for details on the European samples and the note to Table 1 for an explanation of the variance components presented in the table. In the case of the United States, we simply reproduce the figures reported in Table 1 for the 1985–89 to 2010–18 period. The column *Total change (by decade)* reports the annualized change in the variance multiplied by 10. Since countries are observed over different time frames, the transformation is used to make the changes in the variance comparable across countries.

rise of earnings inequality, Europe keeps looking much like the United States in the 1980s and 1990s when inequality growth in total income was almost entirely driven by changes in labor income inequality. This suggests a possibility that the contribution of capital income to inequality in Europe may grow in the years to come if high earners—who relatively benefit from the growth in labor income inequality—start accumulating relatively more wealth and receive more capital income down the road.

Another interesting difference between European countries and the United States is that education does not play quite as large a role in inequality growth on the other side of the Atlantic. This is shown in Table 3, which repeats the decomposition reported in Table 1 for all five countries. For France, the percentage changes are difficult to interpret because they are normalized relative to a modest change, especially in the case of women. In the three other European countries, the between-group component linked to changes in returns to education is smaller than in the United States, and is even negative in the United Kingdom. This finding is consistent with Blundell, Green, and Jin (2016), who find that the returns to education did not change much in the United Kingdom in recent years. As in the United States during the 1980s and 1990s, the most important

component of inequality growth in the three European countries besides France is the within-group component among high-school graduates—that is, the "residual" component unlinked to experience and education factors. Composition effects are also quite large in European countries, reflecting the fact that the workers have grown older and more educated in these countries. Germany is, to some extent, an outlier relative to the other European countries. In particular, education has played a substantial role in the rise of earnings inequality. The within- and between-group components combined explain approximately 27 percent of the increase in income inequality over the German sample period, compared to 12 percent for Italy and 6 percent for the United Kingdom. Importantly, the returns to education have increased substantially, accounting for almost 16 percent of the rise in income inequality. In terms of the role that education plays in the evolution of the earnings distribution, Germany thus lies somewhere between the "average" European country and the United States.

A few key messages arise from the comparison of inequality changes in Europe and the United States. First, the sharp US-European divide in whether income inequality is rising at all, documented by Freeman and Katz (1995), no longer holds in recent data, as inequality in three of the four large European economies has been partly catching up to the higher US level of inequality. Second, and unlike in the United States, capital income is not a significant part of the inequality story in Europe—or at least not yet. This is true without reservation for all four European countries we consider in our analysis. Third, with the slight exception of Germany, education doesn't play as much of a role in inequality growth in Europe, perhaps because the supply of highly educated workers has grown faster in these countries. For example, Blundell, Green, and Jin (2016) discuss this point in the context of the United Kingdom. On the other hand, Germany, as with other trends in labor market outcomes, is becoming more and more the European country that resembles the US experience the most.

Concluding Comments

In an examination of income inequality trends in the United States, the consideration of capital income further accentuates the main story of growing income inequality, as the capital-to-labor income ratio disproportionately increased among high-earning individuals. However, the magnitude of the capital income component is relatively small compared to the predominant source of rising total income inequality, which is labor earnings inequality. Furthermore, various aspects of education—both the gaps between groups defined by education and the growing fraction of highly educated workers—appear to be the predominant force behind the growth in both labor and capital income inequality in the United States, with some role for occupation premia and composition effects linked to the polarization of employment across occupations and space.

Findings for large European economies are more nuanced. While inequality in Germany, Italy, and the United Kingdom grew at least as fast as in the United States in recent years, it remained stable in France. Furthermore, the nature of inequality changes—in particular, the role of capital income and education—was quite different in Europe than in the United States. The modest contribution of capital income to inequality growth in Europe may reflect the fact that US inequality started growing earlier, and has gradually led to an increase in wealth and capital income inequality. Better understanding the nature of the differences in inequality growth in the United States and Europe should be an important priority for future research.

Overall, although the same global forces towards increased inequality appear to be at play in both the United States and Europe, the role of capital income and education in rising inequality remains quite different across countries. Furthermore, the fact that inequality has remained stable in France suggests that country-specific factors can still mitigate other forces pushing toward greater inequality. A detailed investigation of the role of supply, demand, and institutional factors remains essential for understanding similarities and differences in the inequality changes in different countries.

■ *We are grateful to Victoria Angelova and Myera Rashid for expert research assistance, and to Melanie Heiliger for assisting us with the German EVS data, including running and testing our codes.*

References

Acemoglu, Daron, and David H. Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In *Handbook of Labor Economics*, 4th ed., edited by David Card and Orley Ashenfelter, 1043–1172. North Holland: Elsevier.

Acemoglu, Daron, and Pascual Restrepo. 2020. "Robots and Jobs: Evidence from US labor Markets." *Journal of Political Economy* 128 (6).

Alvaredo, Facundo, Anthony B. Atkinson, Thomas Piketty, and Emmanuel Saez. 2013. "The Top 1 Percent in International and Historical Perspective." *Journal of Economic Perspectives* 27 (3): 3–20.

Autor, David H. 2014. "Skills, Education, and the Rise of Earnings Inequality Among the 'Other 99 Percent'." *Science* 344 (6186): 843–51.

Autor, David H. 2019. "Work of the Past, Work of the Future." *American Economic Review Papers and Proceedings* 109: 1–32.

Autor, David H., and David Dorn. 2013. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103 (5): 1553–97.

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

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

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118 (4): 1279–1333.

Autor, David H., David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2020. "The Fall of the Labor Share and the Rise of Superstar Firms." *Quarterly Journal of Economics* 135 (2): 645–709.

Berman, Eli, John Bound, and Zvi Griliches. 1994. "Changes in the Demand for Skilled Labor within U.S. Manufacturing: Evidence from the Annual Survey of Manufactures." *Quarterly Journal of Economics* 109 (2): 367–97.

Blundell, Richard, David A. Green, and Wenchao Jin. 2016. "The UK Wage Premium Puzzle: "How Did a Large Increase in University Graduates Leave the Education Premium Unchanged?" IFS Working Paper W16/01.

Bound, John, and George Johnson. 1992. "Changes in the Structure of Wages in the 1980's: An Evaluation of Alternative Explanations." *American Economic Review* 82 (3): 371–92.

Caines, Colin, Florian Hoffmann, and Gueorgui Kambourov. 2017. "Complex-Task Biased Technological Change and the Labor Market." *Review of Economic Dynamics* 25: 298–319.

Card, David, and Thomas Lemieux. 2001. "Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-Based Analysis." *Quarterly Journal of Economics* 116 (2): 705–46.

Card, David, Jörg Heining, and Patrick Kline. 2013. "Workplace Heterogeneity and the Rise of West German Wage Inequality." *Quarterly Journal of Economics* 128 (3): 967–1015.

Chetverikov, Denis, Bradley Larsen, and Christopher Palmer. 2016. "IV Quantile Regression for Group-Level Treatments, With an Application to the Distributional Effects of Trade." *Econometrica* 84 (2): 809–33.

Daruich, Diego, Sabrina Di Addario, and Raffaele Saggio. 2020. "The Effects of Partial Employment Protection Reforms: Evidence from Italy." Centro Studi Luca D'Agliano Development Studies Working Paper 463.

Devicienti, Francesco, Bernardo Fanfani, and Agata Maida. 2019. "Collective Bargaining and the Evolution of Wage Inequality in Italy." *British Journal of Industrial Relations* 57 (2): 377–407

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach." *Econometrica* 64 (5): 1001–44.

Dustmann, Christian, Bernd Fitzenberger, and Markus Zimmermann. 2018. "Housing Expenditures and Income Inequality." ZEW Discussion Paper 18–048.

Dustmann, Christian, Johannes Ludsteck, and Uta Schönberg. 2009. "Revisiting the German Wage Structure." *Quarterly Journal of Economics* 124 (2): 843–81.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2011. "Occupational Tasks and Changes in the Wage Structure." IZA Discussion Paper 5542.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles and J. Robert Warren. 2000. "Integrated Public Use Microdata Series, Current Population Survey: Version 7.0 [dataset]." Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D030.V7.0. (accessed February 09, 2020).

Fortin, Nicole, and Thomas Lemieux. 2016. "Inequality and Changes in Task Prices: Within and Between Occupation Effects." *Research in Labor Economics* 43: 195–226.

Fortin, Nicole M., Thomas Lemieux, and Neil Lloyd. 2019. "Labor Market Institutions and the Distribution of Wages: The Role of Spillover Effects." Unpublished.

Freeman, Richard B., and Lawrence F. Katz. 1995. *Differences and Changes in Wage Structures*. Chicago: University of Chicago Press.

Goldin, Claudia D., and Lawrence F. Katz. 2007. "Long-Run Changes in the U.S. Wage Structure: Narrowing, Widening, Polarizing." *Brookings Papers on Economic Activity* 38 (2007–2): 135–65.

Goldin, Claudia D., and Lawrence F. Katz. 2008. *The Race between Education and Technology*. Cambridge, MA: Harvard University Press.

Goos, Maarten, Alan Manning, and Anna Salomons. 2014. "Explaining Job Polarization: Routine-Biased Technological Change and Offshoring." *American Economic Review* 104 (8): 2509–26.

Hoffmann, Florian. 2019. "HIP, RIP, and the Robustness of Empirical Earnings Processes." *Quantitative Economics* 10 (3): 1279–1315.

Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce. 1993. "Wage Inequality and the Rise in Returns to Skill." *Journal of Political Economy* 101 (3): 410–42.

Karabarbounis, Loukas, and Brent Neiman, 2014. "The Global Decline of the Labor Share." *Quarterly*

Journal of Economics 129 (1): 61–103.
Katz, Lawrence F., and Kevin M. Murphy. 1992. "Changes in Relative Wages, 1963–1987: Supply and Demand Factors." *Quarterly Journal of Economics* 107 (1): 35–78.
Krueger, Alan B. 1993. "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984–1989." *Quarterly Journal of Economics* 108 (1): 33–60.
Lee, David S. 1999. "Wage Inequality in the United States during the 1980s: Rising Dispersion or Falling Minimum Wage?" *Quarterly Journal of Economics* 114 (3): 977–1023.
Lemieux, Thomas. 2006a. "Postsecondary Education and Increasing Wage Inequality." *American Economic Review* 96 (2): 195–99.
Lemieux, Thomas. 2006b. "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?" *American Economic Review* 96 (3): 461–98.
Luxembourg Income Study (LIS) Database. 2020. "Luxembourg: LIS." http://www.lisdatacenter.org (accessed May 2020).
Machin, Stephen. 2011. "Changes in UK Wage Inequality over the Last Forty Years." In *The Labour Market in Winter*, edited by Paul Gregg and Jonathan Wadsworth, 155-69. Oxford: Oxford University Press.
Manacorda, Marco. 2004. "Can the Scala Mobile Explain the Fall and Rise of Earnings Inequality in Italy? A Semiparametric Analysis, 1977-1993." *Journal of Labor Economics* 22 (3): 585–613
Meyer, Bruce. D., Wallace K.C. Mok, and James X. Sullivan. 2015. "Household Surveys in Crisis." *Journal of Economic Perspectives* 29 (4): 199–226.
Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *Quarterly Journal of Economics* 118 (1): 1–41.
Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2018. "Distributional National Accounts: Methods and Estimates for the United States." *Quarterly Journal of Economics* 133 (2): 553–609.
Rothbaum, Jonathan L. 2015. "Comparing Income Aggregates: How Do the CPS and ACS Match the National Income and Product Accounts, 2007–2012." SEHSD Working Paper 2015-01.
Saez, Emmanuel, and Gabriel Zucman. 2016. "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data." *Quarterly Journal of Economics* 131 (2): 519–78.
Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2019. "Capitalists in the Twenty-First Century." *Quarterly Journal of Economics* 134 (4): 1675–1745.
Steiner, Viktor, and Kersten Wagner. 1996. "Has Earnings Inequality in Germany Changed in the 1980s?" ZEW Discussion Paper 96–32.

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 79–104

An Economist's Guide to Epidemiology Models of Infectious Disease

Christopher Avery, William Bossert, Adam Clark, Glenn Ellison, and Sara Fisher Ellison

Around mid-March 2020, as the United States and much of the rest of the world was facing an unprecedented health threat in the form of COVID-19, an abrupt shift in the tone and policies of the United States and United Kingdom occurred. In early March, Prime Minister Boris Johnson said that "we should all basically just go about our normal daily lives." Likewise, on March 11, President Donald Trump reassured the American people that for "[t]he vast majority of Americans, the risk is very, very low." Just five days later, the Trump administration recommended that "all Americans, including the young and healthy, work to engage in schooling from home when possible. Avoid gathering in groups of more than 10 people. Avoid discretionary travel. And avoid eating and drinking at bars, restaurants, and public food courts" (as reported by Keith 2020). The British government likewise markedly changed course, with a series of partial measures preceding a March 23 lockdown order. Although Trump and Johnson had been receiving briefings about COVID-19 for several weeks, the proximate cause of the

■ *Christopher Avery is Roy E. Larsen Professor of Public Policy, Harvard Kennedy School, Cambridge, Massachusetts. William Bossert is David B. Arnold, Jr., Professor of Science, Emeritus, Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, Massachusetts. Adam Clark is Assistant Professor, Institute of Biology, University of Graz, Graz, Austria. Glenn Ellison is Gregory K. Palm Professor of Economics, Cambridge, Massachusetts. Sara Fisher Ellison is Senior Lecturer in Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Their email addresses are christopher_avery@hks.harvard.edu, bossert@seas.harvard.edu, adam.tclark@gmail.com, gellison@mit.edu, and sellison@mit.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.4.79.

shift in both countries appears to have been the March 16 release of a headline-grabbing epidemiological model produced by London's Imperial College, which predicted that there could be as many as 2,200,000 deaths in the United States and 510,000 in the United Kingdom" (as reported by Landler and Castle 2000).

The Imperial College model was not the only one to feature prominently in public policy. The Institute for Health Metrics and Evaluation (IHME) at the University of Washington released and frequently updated state-level estimates which garnered substantial attention. Its predictions contrasted markedly with (the most extreme) ones from Imperial College. Both sets of predictions turned out to be quite far off in important ways. This fact should not be surprising. There is, unavoidably, much uncertainty about key parameters early in an epidemic. It also takes longer to produce models that use frontier methods and incorporate data from multiple sources. Still, the models can be faulted for providing standard errors that did not accurately reflect the degree of uncertainty underlying the course of the epidemic.

Given the importance of the topic and the impact that these early models had, it is not surprising that many economists quickly became interested in applying their skills to improve understanding of the COVID-19 pandemic. One goal of this paper is to provide an overview of the extant epidemiological literature to facilitate the work of economists who wish to make incremental contributions. We begin by introducing the classic SIR (susceptible/infected/recovered) model, which serves as the basis of much of modern epidemiology of infectious disease, both theoretical and empirical. As we will discuss, the classic model is useful for building intuition about the possible paths of a pandemic. Researchers typically build on this model in a variety of ways, depending on the specific research question, the characteristics of the epidemic, and the available data. We then turn to methods and challenges of implementing these models in empirical epidemiology. With this background in place, we return to the two high-profile forecasting models, explain where they fit into the landscape of empirical epidemiology, discuss the policy imperatives which drove their prominence, and offer critiques. Finally, we consider the related economics papers, ones that expand on SIR-type models, leverage them to provide policy advice, and offer estimates that could help inform them.

The COVID-19 pandemic poses a wealth of policy challenges. We believe that there are fruitful synergies for economists who acquaint themselves with some basic epidemiology models and empirical techniques. We then consider how their economist's toolbox could dovetail with the existing epidemiology literature to produce useful insights.

Epidemiological Theory

Epidemiological theory has been rooted in empirical facts from the start. In 17th-century London, haberdasher turned statistician John Graunt kept weekly records of the causes of death in London parishes. He used these data to estimate the risks of dying from different diseases. His work was instrumental in the

development of biostatistics, demography, and epidemiology. After him, doctors and medical researchers started relying on statistics and then statistical models to help them predict the spread of infectious disease. In the 18th century, Daniel Bernoulli (1766) devised the first true epidemiological model to study the spread of smallpox. In 1906, W.H. Hamer suggested that the spread of infection should depend on the number of susceptible and infected people. He introduced the mass action law for the rate of new infections. Kermack and McKendrick (1927) leveraged these insights to create the SIR model—the workhorse model still the basis of much of modern epidemiology.

In the past century, the field of epidemiology has advanced along lines similar to those of economics. Theorists have developed more sophisticated models to bring out many insights. In recent years the field has taken an empirical turn, developing increasingly sophisticated models that leverage vast and detailed new data sources. It should be noted that just as a relatively small share of economists focus on real-time forecasting of the economy, a relatively small share of epidemiologists focus on real-time forecasting of new pandemics. Epidemiology is a much broader subject, encompassing the study of the distribution and determinants of health and disease outcomes across various populations. The particular niche of the epidemiology literature that is especially relevant for the current pandemic are the models that focus on the spread of an infectious disease. We will start with a discussion of the workhorse model in this class, the SIR model. We note that this classic model both offers basic insights and provides a tractable framework amenable to being built upon.

The Standard SIR Model

SIR is an acronym for the three states (sometimes referenced as "compartments") in the model: Susceptible, Infected, and Recovered. At each time t, each member of the population is in one of these states, with proportions in these states given by $S(t)$, $I(t)$, and $R(t)$ where $S(t) + I(t) + R(t) = 1$ for a population of unit mass.

There are only two ways to move from one state to another. First, currently infected people may become noninfectious and move to the recovered state. Second, a susceptible person can contract the disease through contact with a currently infected person. People in the recovered state may still be sick (or even dead) but they share two key characteristics: they are not infectious and also not susceptible to future infection. Transition rates between states are governed by parameters γ and R_0, which serve as summary statistics for (1) the recovery rate and (2) the number of people an infectious person would infect over the course of their disease in a fully susceptible population.

One way to motivate the model is to suppose that agents are uniformly randomly matched in continuous time. Assume that each meets on average $R_0\gamma$ others per unit time and that any susceptible agent matched with an infected agent becomes infected. As a result, new infections occur at a flow rate of $\gamma R_0 S(t)I(t)$ per unit time. Suppose also that each infectious agent recovers with probability γ per unit time, creating a flow of $\gamma I(t)$ individuals per unit time moving from the Infected to the

Recovered state. These dynamics can be summarized by the following continuous time dynamic equations for the values of $S(t)$, $I(t)$, and $R(t)$ given the two possible transitions from S to I for new infections and from I to R for sick people who become non-infectious:

$$\dot{S}(t) = -S(t)\,I(t)\,R_0\gamma,$$

$$\dot{I}(t) = S(t)\,I(t)\,R_0\gamma - \gamma I(t),$$

$$\dot{R}(t) = \gamma I(t).$$

The number of periods that an infected agent remains in the infected state follows an exponential distribution with parameter γ, so the expected amount of time in the infected state is $\frac{1}{\gamma}$. With $R_0\gamma$ contacts per person per unit time with others, each infected person has an expected number of R_0 contacts while infected. That is, the parameter R_0 can be thought of as the expected number of people that a newly infected person will directly infect in a population where everyone is susceptible.[1]

The initial level of infection at time 0 is another exogenous parameter of the model and is typically assumed to be quite small (for example, one infection per 10 million people). If $R_0 > 1$, the number of infections is larger than the number of recoveries in early periods, while the proportion in the susceptible state remains close to 1. As a heuristic approximation, we would expect contacts with people infectious at time 0 to directly produce a total of $R_0 I(0)\,S(0)$ new infections, which is approximately $R_0 I(0)$ if $S(0)$ is close to 1. This set of new infections would produce approximately $R_0^2 I(0)$ subsequent new infections, and these would produce $R_0^3 I(0)$, and so on. For this reason, the initial growth rate of infections in an SIR model with $R_0 > 1$ is approximately exponential. Formally, one equilibrium of the system is $S(t) = 1$, $I(t) = 0$, $R(t) = 0$ for all t, but if this equilibrium is locally unstable if $R_0 > 1$, then adding a small number of infected agents leads to contagious growth of $I(t)$. By contrast, an equilibrium with $I(t) = 0$ is locally stable if $R_0 < 1$, as a small infection dies out in that case.

Over time, the growth rate of infections declines because the proportion of people in the susceptible state diminishes continuously as the infection spreads. Regardless of when the infection takes place, each infected person has an expected number R_0 of contacts with others while infectious, but as time passes, more and more of those contacts are with people who are not susceptible. The model has a "herd immunity" threshold of $\bar{S} \equiv 1/R_0$. When $S(t) = \bar{S}$, the expected number of people that a newly infected person will directly infect is equal to 1. The important implication of this property is that once the fraction of the population that is

[1] A common alternative description of the SIR model defines $\dot{S}(t) = -S(t)\,I(t)\,\beta$ and $\dot{I}(t) = S(t)\,I(t)\,\beta - \gamma\,I(t)$, and then identifies R_0 separately as the ratio $R_0 = \frac{\beta}{\gamma}$. It is also equivalent to assume a proportionally higher probability $KR_0\gamma dt$ (where K is a known positive constant) that any pair of agents meet in combination with probability $\frac{1}{K}$ that a susceptible agent matched with an infected agent becomes infected.

susceptible is below the herd immunity threshold $\bar{S}$, a small infection introduced into the population will die out with the size of the infectious population never increasing.[2]

Importantly, note that reaching "herd immunity" does not mean that people will not continue to be infected. New infections continue to occur. They are just outnumbered by recoveries that are occurring. When R_0 is large, the number of people who are infectious when the herd immunity threshold is reached is large, so being limited by the number of recoveries is not comforting. Indeed, in these models there can be substantial "overshooting" with many more than $1 - \bar{S}$ people eventually infected. The number of people who escape the epidemic does not have as simple a formula, but is obviously very important practically. In an uncontrolled epidemic, it can be described as the solution to a simple implicit equation.[3] Numerical examples indicate that overshooting can be dramatic with a significant fraction of the population getting infected after herd immunity is reached. For example, with $R_0 = 2$ we reach "herd immunity" when half the population has been infected, but the infection will not completely die out until another 30 percent of the population has been infected. With $R_0 = 2.5$, herd immunity is reached when 60 percent have been infected, but only 11 percent of the population will remain uninfected in an uncontrolled epidemic. In short, even with a moderate R_0, few escape an uncontrolled epidemic. The "social distancing" policies that have been used to suppress COVID-19 infection rates are essentially an attempt to reduce R_0.

One other noteworthy feature of SIR models is that for many values of R_0, the time-path of new infections (and deaths) has a shape that is fairly symmetric about its peak and looks somewhat like a normal density. This provides a potential explanation for one of the earliest empirical observations in epidemiology: Farr (1840) noted that the time series of deaths in a smallpox epidemic and in four other epidemics "which have not yet been effectually controlled by medical science" were roughly symmetric and bell-shaped. Figure 1 below reproduces Figure 1A from Ferguson et al. (2020) illustrating the predictions of their SIR-like model for Great Britain and the United States.

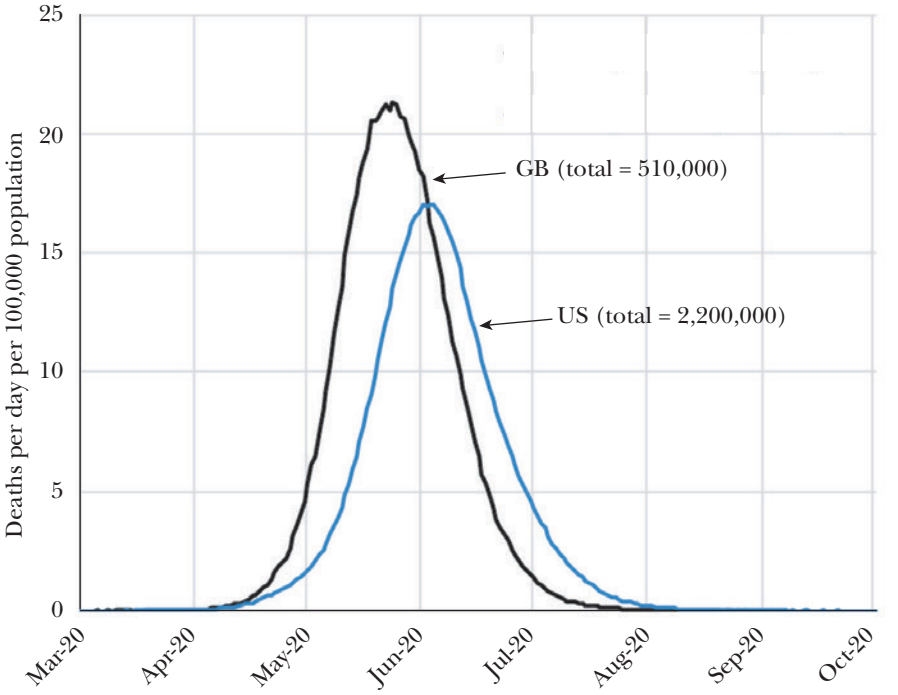
Some Conceptual Lessons from the Standard SIR Model

When a serious contagious disease becomes prevalent, two reactions will typically occur: people will modify their behavior to avoid getting sick and governments will enact policies aimed at slowing or stopping the spread. We can think of the original R_0 as a compound parameter, one that embodies both the underlying biological ability of the pathogen to jump from person to person in various types of

[2] Formally, the herd immunity threshold is such that $S(t) = S$, $I(t) = 0$, $R(t) = 1 - S$ is a stable equilibrium in the model for any $S \leq \bar{S}$.

[3] Formally, we can define the fraction who escape infection, $S(\infty)$, as $S(\infty) \equiv \lim_{t \to \infty} S(t)$. The equation that can be solved to find it is $S(\infty) = e^{-R_0(1-S(\infty))}$. Intuition for the formula is that $1 - S(\infty)$ agents are eventually infected. Each on average has R_0 interactions with others that would cause infection in someone who is susceptible. So the probability of escaping infection is the probability of zero events given a distribution that is Poisson with mean $R_0(1 - \mathrm{S}(\infty))$.

Figure 1
Unmitigated Epidemic Scenarios from Imperial College Model



Source: Figure reproduced from Ferguson et al. (2020)Figure 1A: "Unmitigated epidemic scanarios for GB and the US. (A) Projected deaths per day per 100,000 population in GB and US."

interactions as well as the number of interactions of each type that people have in the ordinary course of their daily lives.[4] As self-interested behavior and government policies reduce interactions, it is as if the R_0 parameter in the equation describing how infections transmit is reduced to some time- and state-dependent variable R_0^t.[5] It is important to remember that all the parameters of SIR models are simple encapsulations of more complex biological events. The cycle of infection involves the population biology of the pathogen outside the host, the behavior and population

[4]This approach has parallels to a classic predator-prey theory in biology, whose models have almost exactly the same form and dynamics as an SIR model. In that literature, there is a parameter governing transition from "freely roaming" to "prey," which is a compound parameter with a fixed attack rate for a particular predator-prey combination as well as a contact rate between predator and prey, which can vary geographically and over time. See Gotelli (2008) for a description.
[5]See Chernozhukov, Kasaha, and Schrimpf (2020) and Goolsbee and Syverson (2020) for empirical evidence on the impact of endogenous behavioral changes and various government policies.

biology of the host, and the interaction of the pathogen and the host. Spatial, temporal, and between-host differences in the details of these events lead to the heterogeneity of the parameters that modelers now find important. While much of epidemiology is focused on understanding these details, they are typically absent from the models currently used to predict the course of diseases.

Policies that reduce the reproduction rate R_0 are often described as "flattening the curve," referring to the graph that shows the rise of cumulative infections over time. A change in behavior that reduces R_0 to R_0^t at any time t affects the fraction of the population that permanently escapes infection. But the standard formula for the herd immunity threshold remains relevant to thinking about the possible long run outcomes: if we are not in the herd immunity region—that is, if $S(t) > \bar{S}$— then the infection will once again spread if government restrictions are removed and people go back to their normal behaviors. If we are in the herd immunity region, then the infection will die out even if all restrictions are removed. Indeed, in this way the SIR model illustrates a clear intuition for how temporary policies can provide long-term benefits: implementing policies that reduce R_0^t at future times when we are approaching the herd immunity threshold will reduce overshooting.

In the case of COVID-19, reaching the herd-immunity threshold is widely believed to entail a devastating loss of life. The SIR model suggests that two other approaches may be appealing in such situations. First, we might put in place policies to reduce R_0^t with the intention of maintaining those policies until a vaccine is developed, thereby keeping the system from ever reaching the herd-immunity region. Second, we might enact more aggressive temporary measures for a period of time sufficient to drive prevalence to a level that is low enough so that less economically costly means of keeping $R_t \equiv R_0^t S(t)$ below 1 become feasible. For example, Hong Kong's suppression of COVID-19 has involved, among other measures, hospitalizing everyone who tests positive to ensure isolation and conducting aggressive contact tracing. This is extremely expensive on a per-infected person basis but has cost trillions less than the US approach, not to mention limiting Hong Kong's loss of life.

The SIR model is also helpful for thinking about vaccines. Vaccines are typically not perfect and neither available to nor willingly received by everyone. Suppose, for instance, that a vaccine was effective in preventing the disease completely and permanently in 60 percent of the people who received it and did nothing for the other 40 percent who received it. Administering such a vaccine to the entire population with, say, 10 percent infected or recovered would result in an additional $0.9 \times 0.6 = 54$ percent of the population immune, so that $S(t) = 1 - 0.1 - 0.54 = 0.36$. Depending on the value of R_0, that number could be sufficient to achieve herd immunity. Achieving herd immunity via a vaccine rather than via infections is also advantageous in that it mitigates overshooting.

Variants of the SIR Model

There are many variants of the SIR model. As usual, the choice to add or subtract complexity from a model should depend on what one is studying. Common variants of the SIR model add additional disease states, referred to as "compartments,"

to provide a more realistic model of disease progression and transmission. The SEIR model includes an "exposed" state to account for individuals who have been infected with the disease but are not yet themselves infectious (Hethcote 2000; Li and Muldowney 1995). The SAIR variant includes an "asymptomatic" compartment for individuals who are infectious but may never develop symptoms. Because of the apparently strong contribution of asymptomatic and pre-symptomatic carriers to the spread of COVID-19, these variants, and particularly the SEIR model, have been quite common in recent epidemiological studies (for example, Kissler et al. 2020; Prem et al. 2020). Epidemiologists sometimes also introduce additional compartments not to reflect disease states but as a mathematical means of making the transmission time process more flexible as in Champredon et al. (2018), although this aim can be accomplished directly as in Zhigljavsky et al. (2020). These variants may be especially useful if one were interested in studying the impact of policies for which timing within the disease cycle is critical, like protocols for testing, contact tracing, and quarantining. For an excellent review of many of these extended forms, see Blackwood and Childs (2018).

A broader category of models divides compartments even further into dozens or even hundreds of different geographic and age states and then allows contact, infection, and recovery rates to vary across classes (Blackwood and Childs 2018; Hethcote 2000). Ebola, for example, is spread through contact with bodily fluids even after death, and one might capture this effect on disease dynamics by considering populations of health care and funeral workers (Champredon et al. 2018). Given the current understanding about how COVID-19 seems to be transmitted, it is easy to think of subpopulations who will have many more risky interactions than average: those living in crowded urban apartments, frequenting bars and nightclubs, using public transportation, attending crowded religious services, working in a nursing home, and so forth.

Models with heterogeneous subpopulations again behave much like the classic SIR model whereby the growth rate of a contagious disease is initially exponential then slows (and eventually dies out) over time (for example, Diekmann, Heesterbeek, and Metz 1990; Dushoff and Levin 1995; Lajmanovich and Yorke 1976). A common pattern in these models is that variations in within-class contact or transmission rates across subgroups produce a faster overall spread of infection than in a well-mixed SIR, with infections concentrated in certain high-risk subgroups. Thereafter, however, dynamics tend to slow down relative to a well-mixed model because contact rates between subgroups are typically lower than the average transmission rate (Bolker 1999). In general, these features tend to lead to less complete spread of diseases in age- and spatially structured models than an analogous homogeneous SIR model, although this is not always the case (Gomes et al. 2020; Hébert-Dufresne et al. 2020). Britton, Ball, and Trapman (2020) provide an illustration in which heterogeneity reduces the herd immunity threshold from 60 to 43 percent. In addition, heterogeneity can also lead to a longer overall persistence of diseases. For example, geographic structure can make it difficult to eradicate a disease fully, allowing periodic resurgences (Lloyd and May 1996).

The polio virus provides an example of the perverse impacts that can emerge from heterogeneity. Changes in hygiene practices in the United States around the middle of the twentieth century led to a decrease in infectiousness in polio, which in turn, led to an increase in its average age of onset. Because younger children typically experienced much milder cases of the virus, this increase in age of onset led to an overall increase in the mortality and morbidity associated with being infected with polio, which persisted until the widespread adoption of a vaccine (Melnick 1990).

Real-world disease states and processes are more complex than those assumed in all of these models, of course. For example, "infected" could be treated as a multidimensional continuum of states, instead of a single state. People can vary in the severity of their symptoms, their health outcomes, and the degree of infectiousness. Likewise, whether an exposure results in an infection can depend on the nature and dosage of the exposure. The extent to which people develop immunity will vary. All of these factors are subject to individual, spatial, and temporal heterogeneity.

Empirical Epidemiology

The field of epidemiology does not divide itself into theory and empirical work as neatly as does economics. There is more diversity in research styles and questions. It does appear, though, that like economics (as discussed in Angrist et al. 2020), the field of epidemiology has become more empirically oriented over time. Most relevant to economists, perhaps, are branches estimating parameters of disease processes, forecasting the courses of epidemics, and estimating policy effects. As noted above, forecasts by epidemiologists of the future course of COVID-19 received tremendous attention in the early days of the epidemic. These forecasts can combine theoretical modeling, calibration of some parameters, and estimation of others. Broadly speaking, forecasting models are often regarded as falling into two main styles. Those based on SIR-type models are in a class called "mechanistic," which, like structural empirical models in economics, assume that a model is exactly correct and calibrate or estimate parameters to obtain a predictive model. There is another class of predictive models termed "phenomenological," which may be motivated by theories of disease spread but are not derived directly from those theories. Instead, they posit a functional form for the evolution of cases or apply time-series methods to predict future outcomes based on available observations. This distinction is not a neat one, however, and forecasts can combine elements of both types.

In economics, choice of empirical model and technique is often driven by realities of data quality and availability. Economists interested in policy evaluation have, for instance, invested enormous effort into developing techniques for causal inference with observational data, which is what economists often have to work with. Something similar is true for epidemiologists interested in forecasts: their models are designed to leverage the data available on an epidemic in its earliest crucial

stages to greatest advantage. These early numbers tend to come from boots-on-the-ground efforts such as contact tracing or case counts, and they can be used to estimate parameters of either phenomenological or mechanistic models. To be clear, data from contact tracing differs from case counts in that it has information about the source of and the resulting infections from a particular infection, but it may not include most or all infections. Case counts attempt to document all infections, but not the tree of connections among them.

Mechanistic Forecasts

Even under ideal circumstances, reliably estimating parameters of mechanistic epidemiological models, such as the SIR, can be quite challenging due to their nonlinear and dynamic nature. The simplest idea for estimating R_0—that is, making a list of initial infections, tracking down the number of additional infections that can be traced directly to each of those initial ones, and dividing to obtain an estimate of R_0—is not an accepted practice due to the fact that incomplete contact tracing and asymptomatic cases would lead to downward-biased estimates. Instead, researchers often employ some more sophisticated variant of the following two-step method-of-moments approach: start with the log growth rate of the epidemic as implied by an SIR model, $\gamma(R_0 - 1)$, and equate that to an empirical log growth rate from the case counts. To identify γ and R_0 separately, then, one can use (potentially incomplete) contact tracing data to infer the distribution of length of time between infections, which helps tie down γ.

Most of us have internalized the notion that more data always lead to better estimates, but a counterintuitive situation can exist here. As the epidemic spreads and more data become available, the quality of (at least some of) the data can be compromised. First, contact tracing efforts will inevitably fall behind in a fast-growing epidemic, and the resulting data might be increasingly lower quality. Second, as an epidemic grows, behavioral responses can emerge, which could contaminate an estimate of R_0. Third, increased testing can identify asymptomatic cases which could contaminate case growth rates, because cases which would not have been included in early case counts are included in later ones. In short, more data can lead to worse estimates, as discussed in Ferretti et al. (2020). There is a trade-off, though: these limited sample sizes early in an epidemic make capturing heterogeneity of many types problematic, to say nothing of capturing changes in parameters over time.

We should stress that epidemiologists have studied these issues in depth for many years. Asymptotic analyses of the properties of maximum likelihood estimation and other estimators of parameters in homogeneous and heterogeneous SIR models can be found in Rida (1991) and Britton (1998). Markov Chain Monte Carlo methods for the Bayesian estimation of heterogeneous SIR models are described in Demiris and O'Neill (2005). And modern applications of disease models typically involve parameterization approaches that are more sophisticated than those described above. For examples of work along these lines, useful starting points include Mills, Robins, and Lipsitch (2004), Massad et al. (2010), and Viboud et al. (2018).

Phenomenological Forecasts

In contrast to mechanistic methods, phenomenological approaches are often relatively straightforward to implement for the early stages of an epidemic. Early case data are used to fit the assumed growth curve (for instance, using maximum likelihood estimation). As additional case data come in, the parameter estimates are refined to reflect the new information. Information on the source of any particular case, typically provided by contact tracing, would not be necessary. With limited early data, it can be difficult to estimate as many parameters as one would want to estimate for a realistic compartmental (mechanistic) model, and this fact can make simple phenomenological approaches appealing. For example, Tuite and Fisman (2018) use a simple functional form with just three parameters, estimated by maximum likelihood, in which the way an epidemic declines is determined by one of the parameters. They note that they "are agnostic about the nature of factors that slow growth, but they could be postulated to include behavioural change, public health interventions, increased immunity in the population, or any other dynamic change that slows disease transmission."

As epidemics progress, phenomenological approaches that use time-series techniques to predict changes remain well-suited to making near-term predictions. These models can be less useful, however, for other tasks. Observation error can rise as larger swaths of a population are infected and contact tracing becomes less reliable, and tightly parameterized models lack the flexibility to respond to qualitative changes in disease behavior that are inconsistent with earlier apparent patterns. For example, a model which posits a symmetric, bell-shaped evolution of cases over time cannot accommodate repeated changes in the rate of spread due to changing regulations, changing public perception, and "quarantine fatigue." In a later section, we will see how early fits from the IHME model accurately characterized initial growth rates in case numbers across much of the United States, but its predictions of peak infection numbers and long-term dynamics have proven much less reliable.

Policies and Causal Inference

Epidemiologists and other health researchers have long been interested in the effects of healthcare interventions. The use of randomized controlled trials—often called the "gold standard" for causal inference—was pioneered by health researchers. During epidemics, however, the earliest data available are typically observational. Even in randomized trials, noncompliance raises concerns about selection biases. And, of course, the very nature of an infectious disease implies that a treatment applied to one agent may affect others. As a result, epidemiologists have recognized that the methods most commonly used in other medical fields for policy evaluation may be less appropriate for epidemiological applications (Halloran and Struchiner 1995; Hernán and Robins 2006). By now, however, epidemiologists have developed a variety of techniques to address field-specific concerns (for an extensive exposition, see Hernán and Robins 2020).

Analyses of Genomic Data

Analysis of the SARS-CoV-2 genome has revealed thousands of different strains of the virus circulating around the world (for the current phylogeny, consult nextstrain.org/ncov/global). The medical community has many reasons to be interested in these multiple strains. For instance, they may differ in communicability or virulence, or there could be less-than-perfect immunity across strains. Korber et al. (2020) present laboratory and epidemiological evidence suggesting that the COVID-19 variant which is now most common is more infectious than the strain that was dominant in Wuhan.

For the purposes of estimating epidemiological models, another immediately useful application of these techniques is to trace the spread of various mutations to determine where and when epidemics began in various regions. In fact, genomic data can serve as a type of substitute for contact tracing or detailed micro-level data on social networks and other human interactions, allowing researchers to trace the source of a particular group of infections without ever knowing anything about the agents' contacts. Researchers in Israel used genomic data, for instance, to produce the often-cited fact that 80 percent of all COVID infections there were caused by 1–10 percent of infected agents (Miller et al. 2020). In another genomic study, Worobey et al. (2020) note that although cases have been reported as early as January 2020 in the United States and Europe, genetic evidence suggests that these introductions failed to spread, and that it was only though later introductions at higher incidence that SARS-CoV-2 was able to establish in the general population. If these findings hold up in follow-up research, they may indicate that even if the virus cannot be fully eradicated, control measures may well prove to be effective if incidences can be brought low enough.

Early High Profile Models—What Went Wrong?

The introduction recounted how an early prediction model from Imperial College had a seemingly huge effect on policy decisions in the United States and the United Kingdom. In fact, one could argue that policy imperatives drove the prominence of that and another high-profile prediction model from IMHE early in the pandemic. Policy-makers were desperate for guidance on mask-wearing and social distancing, predictions on the number of intensive care hospital beds necessary in a particular city, likely timing of peak infections, and so forth. Those two models were up and running early in the pandemic and provided those numbers that policy-makers needed. It is instructive to take a closer look to understand how their predictions were produced and what ultimately went wrong.

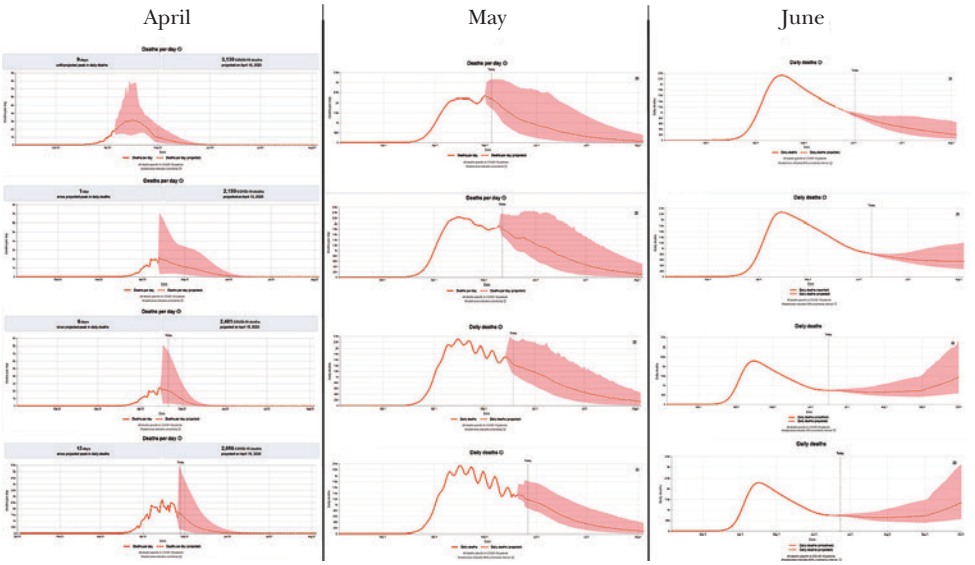
The headline-grabbing figures from the Imperial College model were the most extreme predictions out of many that they produced. They arose from assumptions that governments would not mandate any mitigation strategies, such as mask-wearing or social distancing, and indeed that people would not choose to engage in any of those strategies themselves. Those assumptions were often omitted from the initial

reporting and public discussion of the predictions. Much of the Imperial College report, however, consisted of discussions of the potential impact of such policies, along the lines of an earlier policy discussion on mitigating pandemic influenza in Ferguson et al. (2006). Some information about the details of the Imperial College model were given, but initially the source code was not public. The early reports made certain details clear: the model was based on the familiar SIR framework and that extreme predictions were derived assuming that neither official actions nor individual choices would be taken to slow the spread of the virus. The R_0 term was taken as a single, fixed parameter, with a value of 2.4. Their estimated death rate for those infected was 0.9 percent. Both estimates were based on early experience with COVID-19 in places such as China and Italy but obviously associated with significant uncertainty. The source code for the model was eventually released at the end of April, and researchers were able to reproduce its results from its assumptions by early June (as reported by Chawla 2020). Although this delay is understandable, it was also arguably a contributor to confusion surrounding predictions early in the pandemic.

Meanwhile, as the number of COVID-19 cases was ramping up in the United States, alternative predictions were being offered by IHME at the University of Washington. Their phenomenological model began by assuming a particular functional form for how the number of cases in a locality would rise and then fall over time, with location-specific parameters estimated to fit early case numbers. The model could easily be fit separately to data on each state, and predictions were refined as new data came in. The intention was that local officials could then use these location-specific and daily predictions to plan extra hospital capacity and procure medical equipment, which many of them did. The notion, however, of a common functional form—that is, that the basic shape of increase, peak, and decline of infections would be the same in all locations, from Italy to India, from Wuhan province to Topeka, Kansas—seems to ignore crucial information about how mitigation strategies varied across locations and changed over time. More recent versions of the IHME model have taken an alternative approach, as we discuss in a moment. Roughly speaking, the originally publicized IHME model was assuming a bell shape for the daily deaths and trying to find the parameters governing that bell shape based on the early observations. In a model of this form, once growth has started to slow, there will be limited uncertainty about the size or timing of the peak. Also, the bell-shape symmetry implies that deaths will start falling as rapidly as they grew.

Figure 2 shows a series of screen grabs from the IHME model predicting daily US deaths (from the Internet Archive), at approximately one-week intervals starting in early April 2020. The first four predictions, going down the first column and through the end of April, have several common features resulting from the bell-shape assumptions: the predicted shape of deaths over time is symmetric; the predicted number of deaths goes to zero quickly, around June 1; and the error bands are large in the short run and go to zero around the time that the predicted number of deaths goes to zero. Note that in these first four panels, estimates of the parameters are being updated regularly as new data come in.

Figure 2
Weekly Screenshots of the IHME US Deaths Predictions

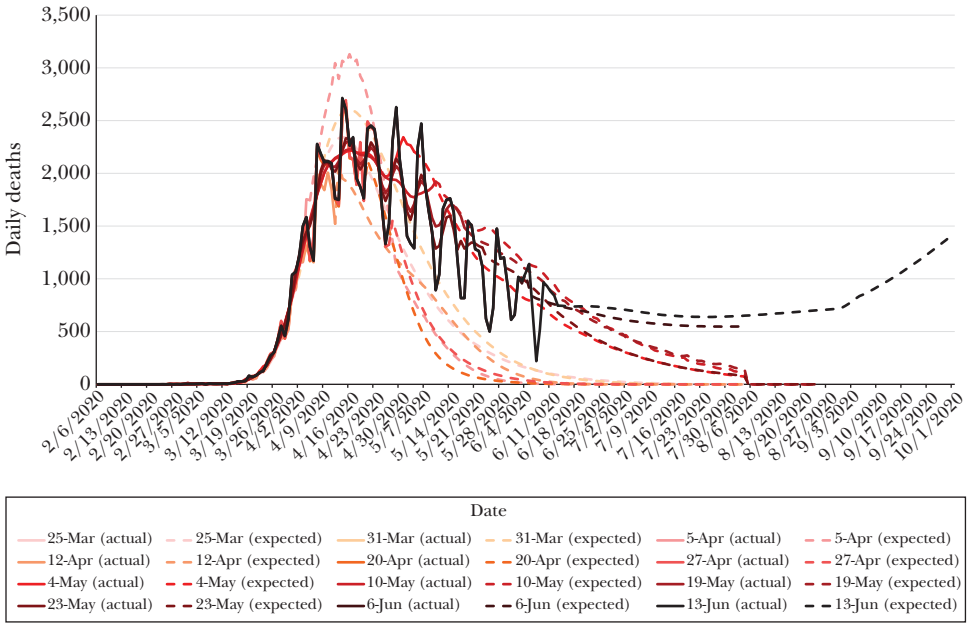


Source: covid19.healthdata.org.
Note: This figure was constructed with a series of screen grabs from covid19.healthdata.org, IHME's website, from the Internet Archive located at archive.org. The screen grabs are at approximate one-week intervals throughout April, May, and June 2020.

In early May 2020, IHME switched away from the curve-fitting approach to a more mechanistic SIR-type framework. The model predicted roughly the deaths in the next few days in a phenomenological way and then fit an SIR-based model to the past and short-term future predictions to generate long-run predictions. The middle column shows that starting in May, the model allowed for asymmetry. It also started using a smoothing algorithm on the existing case data. The way error bands were calculated changed, but error bands still shrunk eventually instead of growing, reflecting that declining deaths implied that epidemics in SIR models in which R_t falls to less than 1 die out in an exponential manner. As a result of these changes, predictions of positive numbers of deaths stretched into summer 2020. Starting in June 2020, the final column of the figure, another substantial change was made to the calculation of error bands, whereby they start small and increase as time proceeds, reflecting increasing, not decreasing, uncertainty in predictions further into the future.[6]

[6]For additional discussion of how the models did not reflect the degree of uncertainty early in the epidemic, see Avery et al. (2020). Stock (2020) also notes the importance of uncertainties that existed early in the pandemic.

Figure 3
IHME US Daily Deaths Predictions Overlaid



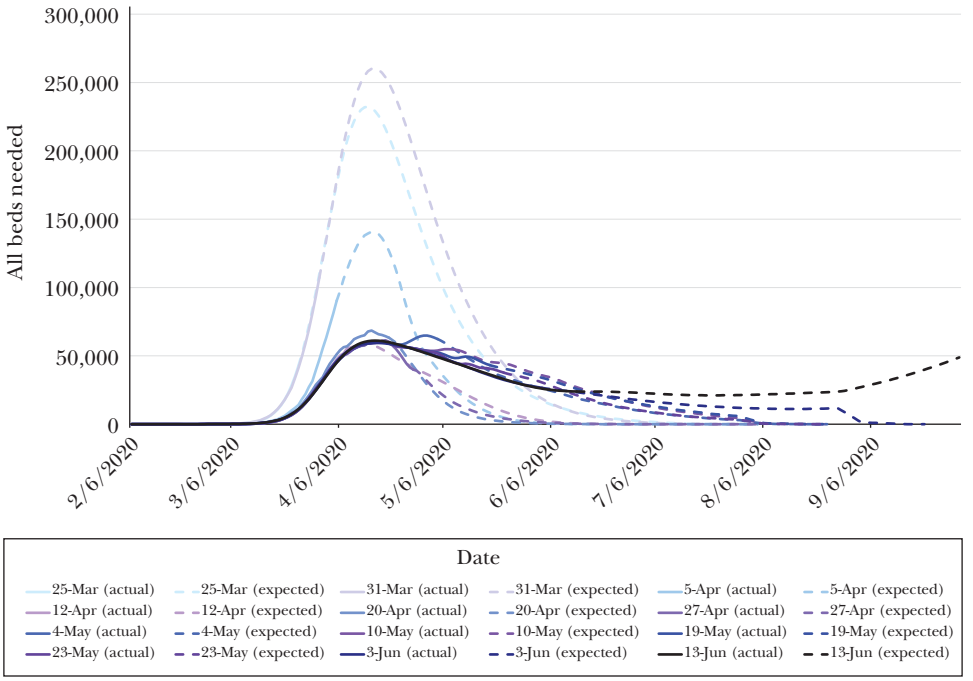
Source: covid19.healthdata.org.
Note: This figure was constructed from data downloaded directly from IHME's website at the University of Washington.

In Figure 3, we overlay these same predictions on a common scale, color-coded so that earlier predictions are lighter. For readability, we do not include error bands. Clearly, the IMHE predictions of US deaths over time change as it becomes clear that the pandemic will not die out at the beginning of the summer, and a symmetric model of US deaths is inaccurate. Even so, the initial predictions of the size and location of the (first) peak were fairly accurate.

Figure 4 shows a different output of the IHME model: predictions of hospital utilization. With this outcome, the initial predictions are starkly different from later ones. Not coincidentally, many locations prepared for much greater hospital utilization during the first "surge" than was needed. We should note that IHME does publish their source code and is forthcoming about changes. That being said, the model is complicated enough that reading through the source code and documented changes is difficult and time-consuming, certainly for us, but also, one would imagine for most researchers.

The Imperial College and IMHE models filled a void early on for policymakers scrambling to understand the pandemic, to decide how strongly to react, to

Figure 4
IHME US Hospital Use Predictions Overlaid



Source: covid19.healthdata.org.
Note: This figure was constructed from data downloaded directly from IHME's website at the University of Washington.

convey policies to constituents, and to allocate resources. But many other predictive models are now available, some with well-designed online dashboards where users can insert different assumptions, some backed by state-of-the-art epidemiology theory, and some leveraging empirical innovations and new information. We cannot hope to survey all of the predictive models here, but both the Centers for Disease Control and Prevention (CDC) and the website FiveThirtyEight.com highlight and compare several of the most well-known and well-received ones.[7] Table 1 shows 15 models highlighted by FiveThirtyEight.com (including IHME), with a few words about their basic approaches and some details about their implementation. These models largely agree in their short-run predictions, but divergence appears at forecasting horizons of six weeks or more. We have organized them by predicted

[7]The Centers for Disease Control has come under criticism from many quarters for allowing political considerations to influence how they present and describe predictive models.

Table 1
Some Predictive Epidemiological Models

Source	*Approach*	*Details*
High Predicted Mortality Level (by Sept. 5th)		
The University of Texas COVID-19 Modeling Consortium, University of Texas https://covid-19.tacc.utexas.edu/projections/	Model 1 uses a curve fitting approach, and Model 2 is an SEIR model with compartment "D" (dead)	Uses anonymized mobile phone data and daily reported deaths to make predictions for three weeks ahead
COVID Scenario Pipeline, Johns Hopkins University https://github.com/HopkinsIDD/COVIDScenarioPipeline	SEIR model	Projects the spread of the epidemic and impacts on health care for different interventions
ERDC SEIR Model, U.S. Army Engineer Research and Development Center https://github.com/reichlab/covid19-forecast-hub/blob/master/data-processed/USACE-ERDC_SEIR/metadata-USACE-ERDC_SEIR.txt	SEIR model with compartments for unrecorded infections and isolated individuals	Uses Bayesian inference to choose parameters
EpiGro, University of Arizona https://www.sciencedirect.com/science/article/pii/S1755436516300329	Curve fitting model	Based on properties of curves implied by SIR model
Medium Predicted Mortality Level (by Sept. 5th)		
DeepCOVID Model, Georgia Tech https://www.cc.gatech.edu/~badityap/covid.html	Deep learning model	Assumes that the effect of interventions is implicitly captured in mobility data
IHME COVID-19 Projections, IHME, University of Washington https://covid19.healthdata.org/united-states-of-america	Hybrid model that incorporates statistical and disease transmission models	Uses social distancing information and mobile phone data to estimate contact between people
COVID-19 Projections Using Machine Learning, Youyang Gu https://covid19-projections.com/	SEIR with machine learning to choose parameters	Estimates incorporate all infected individuals of SARS-CoV-2 virus, not only individuals who tested positive from a COVID-19 test
Columbia University COVID-19 Projections, Shaman Group https://github.com/shaman-lab/COVID-19Projection	Metapopulation SEIR with filtering to determine parameters	Includes projections for daily cases, infections, mortality, and cumulative hospital usage
Global Epidemic and Mobility Model (GLEAM), Northeastern University https://covid19.gleamproject.org/	SEIR model with mobility data	Region-level model with several types of human mobility between regions
Low Predicted Mortality Level (by Sept. 5th)		
COVID-19 Simulator, MGH, Harvard Medical School, Georgia Tech, Boston Medical Center https://www.covid19sim.org/	SEIR model	Includes state-level variations in mobility and tracks hospital usage
Bayesian SEIRD Model, University of Massachusetts https://github.com/dsheldon/covid	SEIR model with additional compartments "D" (death) and "H" (hospitalized-and-will-die)	Employs Bayesian inference and time-varying dynamics
UCLA-SuEIR Model, UCLA https://covid19.uclaml.org/	SuEIR model	Has compartment for unobserved infections
A Shiny App, Iowa State http://www.covid19dashboard.us/	New spatiotemporal epidemic modeling (STEM) framework	Nonparametric model emphasizing 7-day forward projections down to county level
DELPHI Epidemiological Model, MIT https://www.covidanalytics.io/	SEIR with under-detection, hospitalization, and government interventions	Varies effective contact rate and societal/government response by state
LANL Model, Los Alamos https://covid-19.bsvgateway.org/	Dynamic model that forecasts future cases and deaths	Allows for a variety of interventions, resulting in a wide prediction interval

mortality levels. In part, this divergence may reflect different assumptions about how social distancing and government policies will evolve.

As this article was being completed in late summer 2020, it seemed that predictive models about the future of the epidemic had faded from popular discourse. Discussions of reported cases, deaths, and trends seemed, by mid-July, to be getting more attention than forecasts from epidemiological models. Google Trends indicates that searches for "IHME Model" peaked in mid-April and had fallen by 90 percent by early July. Attention by academics also seems to have fallen: Google Scholar indicates that Ferguson et al. (2020), released on March 16, had already been cited 828 times in early July, while the later May 21 report by the Imperial group (Unwin et al. 2020) providing more sophisticated estimates of R_0 for US states had been cited just three times.

One likely reason for the initial surge and subsequent decline of interest in predictive models is that earlier in 2020 they were seen as relevant to policy choices: whether to require businesses to close and people to stay home and how much to invest in hospital bed capacity. By contrast, predictive models appear to have much less relevance to the pressing decisions of fall 2020 such as when to reopen in-person schools. In addition, predictive models have likely lost popular credibility. The initial IHME forecast predicted that the epidemic would all but die out in the United States by early June. The Imperial College model was often linked to its most extreme predictions. Finally, the waning interest may also reflect that the future course of the disease is not readily predictable by any model, but rather, will depend to a considerable extent on how individuals behave and what policies are enacted.

Epidemiology-Related Research in Economics

Economists have responded enthusiastically to demands for COVID-related research and analysis. We cannot attempt to cover this burgeoning literature in its entirety. Rather, our focus will be tighter: on research that leverages SIR-type models, expands upon them, or offers estimates that could help inform them. We chose this sub-literature as our focus because we feel that it is an area where cross-discipline knowledge and the use of complementary models and tools have already continued and will continue to yield real insights.

It is useful to organize much of this sub-literature into three strands. These strands represent salient features of this pandemic as opposed to previous ones, and we feel that economists are well-positioned to make contributions in those three areas. First, economists have recognized the potential endogeneity of parameters such as R_0^t, as the precautions taken could be a function of disease prevalence or current cases. Second, several economics papers have focused on the effects of allowing various types of heterogeneity in SIR-type models. Third, economists have taken the political economy issues involving endogenous social distancing and government policies seriously—issues which could also greatly influence the pattern of R_t over time. We will discuss each of the three strands in turn.

Endogeneity

The R_0^t parameter in an SIR model is a potentially endogenous parameter, which reflects both how easily communicable a particular pathogen is as well as how people behave and interact given the current state of an epidemic It is natural that economists would recognize this endogeneity and model it theoretically and allow for it in empirical analyses. Applying traditional economics approaches to incorporating behavioral responses into epidemiological models is not new and dates back at least to work on the AIDS epidemic in the 1990s (Kremer 1996; Philipson and Posner 1993). Recently, a strand of COVID-related literature accommodating and studying an endogenous reproduction number has emerged. Toxvaerd (2020) and Kudlyak, Smith, and Wilson (2020) develop models that endogenize the social distancing as reflecting a cost and benefit of avoiding infection and discuss impacts on the time path of infections. Farboodi, Jarosch, and Shimer (2020) develop a tractable model of forward-looking individual distancing in which they can compare equilibrium and social optimizing distancing. They calibrate to epidemiological estimates of R_0 from early in the pandemic. They then show that, given a particular choice for the disutility of social distancing, the laissez-faire equilibrium, where social distancing is the result of endogenous individual choices, roughly matches the degree of distancing in the United States as measured by cell-phone mobility data. They find that the optimal government policy in the United States, taking externalities into account, is immediate—but not particularly restrictive—social distancing of long duration. Eichenbaum, Rebelo, and Trabandt (2020) develop another model in which the primary channel for distancing is to reduce consumption of social goods, which is restrictive as a model of distancing activities but creates clean connections to macroeconomic activity.

Goolsbee and Syverson (2020) study endogenous social distancing from an empirical perspective. They provide an estimate of how important endogenous individual actions are relative to government policies designed to lower R_t. Using county-level mobility data in a border discontinuity design, they find that of the 60 percent decrease in US activity observed, only about 7 percentage points can be explained by government regulations across different states and municipalities. Their research suggests that ignoring endogeneity in these models could be problematic and could, in particular, lead researchers to mistakenly attribute effects on disease dynamics to government policies. Chernozhukov, Kasaha, and Schrimpf (2020) find substantial causal effects of government policies using a more sophisticated dynamic model of consumer choices, while still finding that providing information on risks is also quite important.

The endogeneity of R_0^t has also been recognized and addressed by epidemiologists. Reluga (2010) is most similar to how some economists have set up the problem—it develops a differential game version of the SIR model in which agents can, at each instant, take a costly social distancing action that reduces their instantaneous probability of infection. It computes equilibria for several sets of parameter values covering scenarios in which the disease spreads at different rates and a vaccine is closer or farther off, and compares equilibrium payoffs to the social optimum.

Reluga (2010) also provides references to earlier literature, much of which is less utility-focused. A recent example of work of this style is Eksin, Paarporn, and Weitz (2019), which discusses variants of the SIR model that make how people distance in response to current or cumulative cases as a primitive (instead of deriving this from a utility function) and notes that distancing could make the long-run fraction infected much lower than would be predicted by an SIR model calibrated in early stages of the epidemic. While economists' first inclination will be to regard it as a drawback that distancing behavior is a primitive rather than derived from dynamic optimization given an assumed utility function, a skeptic could easily note that there is quite limited evidence on the utility-consistency of the ways in which people socially distance over the course of an epidemic, and that models with utility functions calibrated to rationalize how people have distanced in past epidemics may not provide better predictions than would models in which behavior itself is calibrated to behavior in past epidemics.

Heterogeneity

In many branches of economics, it has become standard to incorporate heterogeneous consumer preferences and/or firm profit functions. Given this norm, it is not surprising that economists are also increasingly incorporating heterogeneity into their COVID-related work.

One of the most striking features of COVID-19 is how fatality rates vary with age. The calibrations in Ferguson et al. (2020), for example, assume an infection fatality rate of 9.3 percent for those over 80, 2.2 percent for those 60–69, 0.15 percent for those 40–49, and 0.03 percent for those 20–29. Economic activities also vary with age, of course. Therefore, it is natural to assess the potentially disparate impact that policies may have on different age groups, consider explicitly age-varying policies, or both.

Several recent papers use calibrated multi-population SIR models where subpopulations are interpreted as age groups to discuss the economic and health consequences of lockdown and reopening policies. Rampini (2020) considers a two-population model calibrated to reflect those under and over age 55 and notes that a two-phase reopening in which the young are released before the old can reduce hospital overcrowding, mortality, and economic losses. Favero, Ichino, and Rustichini (2020) and Baqaee et al. (2020) make finer distinctions of subpopulations. The former considers a 15-population model corresponding to subsets defined by five age groups and three occupation types. The latter uses a five population model corresponding to age groups but calibrates interactions between age groups using contact survey data,[8] data on activity differences across occupations, and industry-specific worker age distributions. In other words, they take an estimate of the average R_0 from the epidemiology literature and choose a matrix of subgroup-to-subgroup infection rates that is consistent both with that R_0 and with

[8] "Contact surveys" are distinct from contact tracing. The former simply obtains data on typical daily contacts of randomly-selected people, both within and across various subgroups.

the differences across groups in the contact surveys and mobility data. The results of Baqaee et al. (2020) are sobering: even slow reopening policies that prioritize industries on a GDP-to-risk basis tend to produce conditions that require subsequent reversals of policy with new shutdowns if individuals relax their levels of social distancing. Acemoglu et al. (2020) analyze a much broader class of time- and age-varying policies and provide estimates of the Pareto frontier of optimal policies that minimize economic losses and deaths. They note that age-dependent policies can provide substantial gains relative to uniform policies, with the greatest improvement coming from doing as much as one can to protect those in the oldest group when prevalence is high among those in younger age groups.

Ellison (2020) builds on models in the epidemiology literature that take a broader view of heterogeneity—reflecting that those who ride public transportation or frequent bars will have many more contacts than others in their age group, for example—and discusses their implications for an analysis of COVID-19. Jackson and López-Pintado (2013) is an example within economics. One cautionary observation is that these models have more parameters that need to be calibrated, and long run outcomes can be sensitive to activity levels of the less active, particularly when we are considering relaxing restrictions. It is difficult to calibrate these parameters early in an epidemic, and predictions that do not allow for heterogeneity may be overconfident.

Ellison (2020) also notes that conclusions drawn from applying homogeneous SIR models to a world that is more like a heterogeneous SIR model would be biased in a number of ways. As noted earlier, homogeneous SIR models may substantially overstate the fraction of the population that must be infected in order to achieve herd immunity. A related observation is that (targeted) lockdown policies can also be more cost-effective in heterogeneous populations. There can be substantial gains either from taking permanent measures to reduce spread among the highly active or from temporarily locking down less active groups to minimize overshooting of herd immunity thresholds. We look forward to seeing such heterogeneities incorporated into more policy analyses.[9]

Political Economy

An extraordinary characteristic of this health crisis in the United States is the degree to which it has been politicized, even to the extent that simple precautions like wearing a mask have become freighted with political meaning. Evidence suggests that social distancing and mask-wearing are very important weapons in combating COVID-19 (Abaluck et al. 2020; Chernozhukov, Kasaha, and Schrimpf 2020), so understanding political obstacles to improving, or simply variation in, these behaviors is quite important. A trio of papers attempt to address this issue by looking specifically at the role of the media. They have found evidence of correlation or

[9] Given the substantial fraction of deaths which have occurred in nursing homes, one such extension that seems very natural would be to incorporate a nursing home sector. This would allow one to model impacts of policies like those discussed in Chen, Chevalier, and Long (2020).

causal effects of media consumption on knowledge about COVID-19 and behavior regarding it. Jamieson and Albarracín (2020) find that, controlling for party affiliation and other demographics, use of conservative media was associated with significantly lower levels of knowledge about the virus and the disease characteristics associated with it. Simonov et al. (2020) exploit quasi-random assignments of channel positions in a cable lineup to estimate the effect of full Fox News viewership on non-compliance with stay-at-home orders, finding an increase of 12–25 percent noncompliance. Finally, Bursztyn et al. (2020), also interested in the effect of Fox viewership, exploit a different instrument, the broadcast time of Hannity and Tucker Carlson Tonight relative to sunset in a particular location. They document a much different tone to the COVID-related content on the two shows early in the epidemic and find that areas with greater exposure to Hannity—more dismissive of the risks—experienced significantly more cases and deaths.

Barrios and Hochberg (2020) use data on internet searches to document that Republican-dominated areas perceive less risk from the virus than do Democratic-dominated areas. Finally, Ajzenman, Cavalcanti, and Da Mata (2020) find similar political effects in Brazil, another country struggling with high caseloads and deaths and with a president dismissive of the severity of the pandemic. They find differential effects on behavior following presidential speeches disparaging social distancing, based on the level of political support for the president by location. Additional papers documenting the political divide and its effects on behavior and health outcomes during the pandemic are cited in these papers as well.

Although none of these papers use epidemiological models or methods, their estimates are useful for understanding how the parameters in the epidemiological models might vary over time and by geographic location. In fact, their specifications and results suggest ways in which R_0^t might be parameterized in an empirical model with a variety in covariates.

Conclusion

A symbiotic relationship between academic research and government policy-making existed long before the spring of 2020. Many researchers aim to produce research that is topical, useful, and policy-relevant. In turn, policymakers seek out expert advice and prediction, often in the form of theoretical or empirical models. Our current crisis, however, has put the structure and the mechanics of this relationship in stark relief.

We think that it is important to draw a distinction between two roles that models have served during the pandemic. Models can help us predict, and they can help us understand, and policy-makers have demanded both types. For instance, models can help us predict timing and magnitude of infections and hospitalizations as well as the need for equipment and other resources. The ability to generate detailed predictions for specific localities is important, especially for local decision-makers who have to set policy and allocate resources. Ultimately, though, the test

of the usefulness of these models is typically empirical in nature, whether that be using retrospective data to judge various models after the fact or using previous and contemporary data from similar settings. The opacity of such models may not be entirely unimportant, but it could be second-order: as long as a "black box" works, we may not care what is in it.

Alternatively, models can help us understand. They can help us understand, for instance, an important interaction of factors, or a mechanism which can indirectly affect the spread of a disease. Such models need not be able to generate location- and day-specific predictions of the number of hospital beds needed, but they are no less important in informing policy-making and resource allocation in different ways.

Understanding the process by which these models' predictions and insights can be accessed by policymakers has also gained importance. The normal process of writing, vetting, and publishing scientific and economic research is being stretched to its limits given the urgency of the pandemic. Direct and wide dissemination can work for certain types of knowledge: detailed predictions from empirical models lend themselves to the now ubiquitous COVID "dashboards" that make those predictions available to policy-makers and others with just a click or two. There is no reason to believe that the models which have the best designed websites and interfaces are the ones producing the most careful and accurate predictions, though. Conveying more subtle insights, such as how government policies might interact with endogenous social distancing, seems substantially more difficult but no less important. One would hope that robust lines of communication and established respectful relationships between experts and policy-makers could facilitate such dialogues.

We wrote this paper in hopes of spurring interesting and important research by economists on epidemics and COVID-19, in particular. If this extraordinary period in time also spurs a rethinking of the complicated relationship between research and policy-making, the dialog between experts and non-experts, and the practical uses of both theoretical and empirical modeling, we will all reap the benefits.

■ *We are grateful to Marcy Alsan, Ben Bolker, Amitabh Chandra, Bill Clark, Jonathan Dushoff, Michael Kremer, Nolan Miller, Ziad Obermeyer, Elizabeth Rourke, Bruce Sacerdote, Doug Staiger, Jim Stock, and Richard Zeckhauser for useful conversations and Eva Demsky for outstanding research assistance.*

References

Abaluck, Jason, Judith A. Chevalier, Nicholas A. Christakis, Howard Paul Forman, Edward H. Kaplan, Albert Ko, and Sten H. Vermund. 2020. "The Case for Universal Cloth Mask Adoption and Policies to Increase the Supply of Medical Masks for Health Workers." *Covid Economics* 5 (2020): 147–59.

Acemoglu, Daron, Victor Chernozhukov, Iván Werning, and Michael D. Whinston. 2020 "Optimal Targeted Lockdowns in a Multi-Group SIR Model." NBER Working Paper 27102.

Ajzenman, Nicolas, Tiago Cavalcanti, and Daniel Da Mata. 2020. "More Than Words: Leaders Speech and Risky Behavior during a Pandemic." Unpublished.

Angrist, Joshua, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. 2017. "Economics Research Evolves: Fields and Styles." *American Economic Review Papers and Proceedings* 107 (5): 293–97.

Avery, Christopher, William Bossert, Adam Thomas Clark, Glenn Ellison, and Sara Fisher Ellison. 2020. "Policy Implications of Models of the Spread of Coronavirus: Perspectives and Opportunities for Economists." *Covid Economics* 12: 21–68.

Baqaee, David, Emmanuel Farhi, Michael J. Mina, and James H. Stock. 2020. "Reopening Scenarios." NBER Working Paper 27244.

Barrios, John M., and Yael V. Hochberg. 2020. "Risk Perception through the Lens of Politics in the Time of the COVID-19 Pandemic." Becker Friedman Institute Working Paper 2020-32.

Bernoulli, Daniel. 1766. "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole." In *Mémoires de Mathématique et de Physique*, edited by Académie royale sciences.

Blackwood, Julie C., and Lauren M. Childs. 2018. "An Introduction to Compartmental Modeling for the Budding Infectious Disease Modeler." *Letters in Biomathematics* 5 (1): 195–221.

Bolker, Benjamin M. 1999. "Analytic Models for the Patchy Spread of Plant Disease." *Bulletin of Mathematical Biology* 61 (849).

Britton, T. 1998. "Estimation in Multitype Epidemics." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60 (4): 663–79.

Britton, Tom, Frank Ball, and Pieter Trapman. 2020. "A Mathematical Model Reveals the Influence of Population Heterogeneity on Herd Immunity to SARS-CoV-2." *Science* 369 (6505): 846–49.

Bursztyn, Leonardo, Aakaash Rao, Christopher P. Roth, and David H. Yanagizawa-Drott. 2020. "Misinformation during a Pandemic." NBER Working Paper 27417.

Champredon, David, Michael Li, Benjamin M. Bolker, and Jonathan Dushoff. 2018. "Two Approaches to Forecast Ebola Synthetic Epidemics." *Epidemics* 22: 36–42.

Chawla, Dalmeet Singh. 2020. "Influential Pandemic Simulation Verified by Code Checkers." *Nature,* June 18. https://media.nature.com/original/magazine-assets/d41586-020-01685-y/d41586-020-01685-y.pdf.

Chen, M. Keith, Judith A. Chevalier, and Elisa F. Long. 2020. "Nursing Home Staff Networks and COVID-19." NBER Working Paper 27608.

Chernozhukov, Victor, Hiroyuki Kasaha, and Paul Schrimpf. 2020. "Causal Impact of Masks, Policies, Behavior on Early Covid-19 Pandemic in the US." *arXiv.*

Demiris, Nikolaos, and Philip D. O'Neill. 2005. "Bayesian Inference for Stochastic Multitype Epidemics in Structured Populations via Random Graphs." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (5): 731–45.

Diekmann, Odo, Johan Andre Peter Heesterbeek, and Johan A.J. Metz. 1990. "On the Definition and the Computation of the Basic Reproduction Ratio R 0 in Models for Infectious Diseases in Heterogeneous Populations." *Journal of Mathematical Biology* 28 (4): 365–82.

Dushoff, Jonathan, and Simon Levin. 1995. "The Effects of Population Heterogeneity on Disease Invasion." *Mathematical Biosciences* 12 (1–2): 25–40.

Eichenbaum, Martin S., Sergio Rebelo, and Mathias Trabandt. 2020. "The Macroeconomics of Epidemics." NBER Working Paper 26882.

Eksin, Ceyhun, Keith Paarporn, and Joshua S. Weitz. 2019. "Systematic Biases in Disease Forecasting–The Role of Behavior Change." *Epidemiology* 27: 96–105.

Ellison, Glenn. 2020. "Implications of Heterogeneous SIR Models for Analyses of COVID-19." NBER Working Paper 27373.

Farboodi, Maryam, Gregor Jarosch, and Robert Shimer. 2020. "Internal and External Effects of Social Distancing in a Pandemic." NBER Working Paper 27059.

Farr, William. 1840. "Causes of Death in England and Wales." In *Second Annual Report of the Registrar General of Births, Deaths and Marriages in England*, 100–53. London: W. Clowes and Sons.

Favero, Carlo A., Andrea Ichino, and Aldo Rustichini. 2020. "Restarting the Economy while Saving Lives under Covid-19." CEPR Discussion Paper DP14664.

Ferguson, Neil M., Derek A.T. Cummings, Christophe Fraser, James C. Cajka, Philip C. Cooley, and Donald S. Burke. 2006. "Strategies for Mitigating and Influenza Pandemic." *Nature* 442: 448–52.

Ferguson, Neil M., Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, et al. 2020. *Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand*. Swindon, United Kingdom: Medical Research Council: The Royal Society.

Ferretti, Luca, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. 2020. "Quantifying SARS-CoV-2 Transmission Suggests Epidemic Control with Digital Contact Tracing." *Science* 368 (6491).

Gomes, M. Gabriela M., Rodrigo M. Corder, Jessica G. King, Kate E. Langwig, Caetano Souto-Maior, Jorge Carneiro, Guilherme Gonçalves, Carlos Penha-Goncalves, Marcelo U. Ferreira and Ricardo Aguas. 2020. "Individual Variation in Susceptibility or Exposure to SARS-CoV-2 Lowers the Herd Immunity Threshold." https://doi.org/10.1101/2020.04.27.20081893.

Goolsbee, Austan, and Chad Syverson. 2020. "Fear, Lockdown, and Diversion: Comparing Drivers of Pandemic Economic Decline 2020." Becker-Friedman Working Paper 2020-80.

Gotelli, Nicholas J. 2008. *A Primer on Ecology*. Oxford: Oxford University Press.

Halloran, M. Elizabeth, and Claudio J. Struchiner. 1995. "Causal Inference in Infectious Diseases." *Epidemiology* 6 (2): 142–51.

Hébert-Dufresne, Laurent, Benjamin M. Althouse, Samuel V. Scarpino, and Antoine Allard. 2020. "Beyond R0: Heterogeneity in Secondary Infections and Probabilistic Epidemic Forecasting." https://doi.org/10.1101/2020.02.10.20021725.

Hernán, Miguel A., and James M. Robins. 2006. "Instruments for Causal Inference: An Epidemiologists Dream?" *Epidemiology* 17 (4): 360–72.

Hernán, Miguel A., and James M. Robins. 2020. *Causal Inference: What If*. Boca Raton, FL: CRC Press.

Hethcote, Herbert W. 2000. "The Mathematics of Infectious Diseases." *SIAM Review* 42 (4): 599–653.

Jackson, Matthew O., and Dunia López-Pintado. 2013. "Diffusion and Contagion in Networks with Heterogeneous Agents and Homophily." *Network Science* 1 (1): 49–67.

Jamieson, Kathleen Hall, and Dolores Albarracín. 2020. "The Relation between Media Consumption and Misinformation at the Oustet of the SARS-CoV-2 Pandemic in the US." *The Harvard Kennedy School Misinformation Review* https://doi.org/10.37016/mr-2020-012.

Keith, Tamara. 2020. "Timeline: What Trump Has Said and Done about the Coronavirus." *NPR,* April 21. https://www.npr.org/202%4/21/837348551/timeline-what-trump-has-said-and-done-about-the-coronavirus.

Keppo, Jussi, Elena Quercioli, Kudlyak, Marianna, Lones Smith, and Andrea Wilson. 2020. "For Whom the Bell Tolls: Avoidance Behavior at Breakout in COVID19." Virtual Macro Seminar, 1:25:07. April 2020.

Kermack, William Ogilvy, and Anderson G. McKendrick. 1927. "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society of London A* 115 (772): 700–21.

Kissler, Stephen M., Christine Tedijanto, Edward Goldstein, Yonatan H. Grad, and Marc Lipsitch. 2020. "Projecting the Transmission Dynamics of SARS-CoV-2 through the Postpandemic Period." *Science* 368 (6493): 860–68.

Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, et al. 2020. "Tracking Changes in the SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus." *Cell* 182 (4): 812–27.

Kremer, Michael. 1996. "Integrating Behavioral Choice into Epidemiological Models of AIDS." *Quarterly Journal of Economics* 111 (2): 549–73.

Lajmanovich, Ana, and James A. Yorke. 1976. "A Deterministic Model for Gonorrhea in a Nonhomogenous Population." *Mathematical Biosciences* 28 (3–4): 221–36.

Landler, Mark, and Stephen Castle. 2020. "Behind the Virus Report that Jarred the U.S. and the U.K. to Action." *New York Times,* March 17. https://www.nytimes.com/2020/03/17/world/europe/coronavirus-imperial-college-johnson.html.

Li, Michael Y., and James S. Muldowney. 1995. "Global Stability for the SEIR Model in Epidemiology." *Mathematical Biosciences* 125 (2): 155–64.

Lloyd, Alun L., and Robert M. May. 1996. "Spatial Heterogeneity in Epidemic Models." *Journal of Theoretical Biology* 179 (1): 1–11.

Massad, E., F.A.B. Coutinho, M.N. Burattini, and M. Amaku. 2010. "Estimation of R0 from the Initial Phase of an Outbreak of a Vector-Borne Infection." *Tropical medicine & International Health* 15 (1): 120–26.

Melnick, J.L. 1990. "Poliomyelitis." In *Tropical and Geographical Medicine*, edited by Kenneth S. Warren and Adel A.F. Mahmoud, 558–76. New York: McGraw Hill.

Miller, Danielle, Michael A. Martin, Noam Harel, Talia Kustin, Omer Tirosh, Moran Meir, Nadav Sorek, et al. 2020. "Full Genome Viral Sequences Inform Patterns of SARS-CoV-2 Spread into and within Israel." https://www.medrxiv.org/content/10.1101/2020.05.21.20104521v1.full.pdf.

Mills, Christina E., James M. Robins, and Marc Lipsitch. 2004. "Transmissibility of 1918 Pandemic Influenza." *Nature* 432: 904–06.

Philipson, Tomas J., and Richard A. Posner. 1993. *Private Choices and Public Health: The AIDS Epidemic in an Economic Perspective*. Cambridge, MA: Harvard University Press.

Prem, Kiesha, Yang Liu, Timothy W. Russell, Adam J. Kucharski, Rosalind M. Eggo, Nicholas Davies, Stefan Flasche, et al. 2020. "The Effect of Control Strategies to Reduce Social Mixing on Epidemic in Wuhan, China: A Modelling Study." *The Lancet Public Health* 5 (5): 261–70.

Rampini, Adriano A. 2020. "Sequential Lifting of Covid-19 Interventions with Population Heterogeneity." NBER Working Paper 27063.

Reluga, Timothy C. 2010. "Game Theory of Social Distancing in Response to an Epidemic." *PLOS Computational Biology* 6 (5): 1–9.

Rida, Wasima N. 1991. "Asymptotic Properties of Some Estimators for the Infection Rate in the General Stochastic Epidemic Model." *Journal of the Royal Statistical Society: Series B* 53 (1): 269–83.

Simonov, Andrey, Szymon K. Sacher, Jean-Pierre H. Dubé, and Shirsho Biswas. 2020. "The Persuasive Effect of Fox News: Non-compliance with Social Distancing during the COVID-19 Pandemic." NBER Working Paper 27237.

Stock, James H. 2020. "Data Gaps and the Policy Response to the Novel Coronavirus." NBER Working Paper 26902.

Toxvaerd, Flavio. 2020. "Equilibrium Social Distancing." Cambridge-INET Working Paper Series 2020/08.

Tuite, Ashleigh R., and David N. Fisman. 2018. "The IDEA Model: A Single Equation Approach to the Ebola Forecasting Challenge." *Epidemics* 22: 71–77.

Unwin, H. Juliette T., Swapnil Mishra, Valerie C. Bradley, Axel Gandy, Thomas A. Mellan, Helen Coupland, Jonathan Ish-Horowicz, et al. 2020. "State-Level Tracking of COVID-19 in the United States." https://www.medrxiv.org/content/10.1101/2020.07.13.20152355v1.full.pdf.

Viboud, Cécile, Kaiyuan Sun, Robert Gaffey, Marco Ajelli, Laura Fumanelli, Stefano Merler, Qian Zhang, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani. 2018. "The RAPIDD Ebola Forecasting Challenge: Synthesis and Lessons Learnt." *Epidemics* 22: 13–21.

Worobey, Michael, Jonathan Pekar, Brendan B. Larsen, Martha I. Nelson, Verity Hill, Jeffrey B. Joy, Andrew Rambaut, Marc A Suchard, Joel O. Wertheim, and Philippe Lemey. 2020. "The Emergence of SARS-CoV-2 in Europe and the US." Unpublished.

Zhigljavsky, Anatoly, Ivan Fesenko, Henry Wynn, Kobi Kremnitzer, Jack Noonan, Jonathan Gillard, and Roger Whitaker. 2020. "A Prototype for Decision Support Tool to Help Decision-Makers with the Strategy of Handling the COVID-19 UK Epidemic." https://doi.org/10.1101/2020.04.24.20077818.

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 105–120

Epidemiology's Time of Need: COVID-19 Calls for Epidemic-Related Economics

Eleanor J. Murray

The COVID-19 pandemic is a global and systemic human catastrophe on a scale that hasn't been seen for infectious disease since the 1918 Spanish Flu. The rapid spread of this pandemic has upended daily life more quickly and more widely than any other large modern pandemic, including HIV/AIDS, and has propelled the relatively unknown field of epidemiology into the public spotlight. Prior to 2020, it was a long-standing joke that upon telling someone you were an epidemiologist, you would immediately be asked to provide advice on a mole or skin lesion ("epidem-" being mistakenly associated with "epidermis"). This has shifted dramatically in 2020 with epidemiologists now being asked, "when will COVID end?" But this question also misunderstands the nature of epidemiology.

Epidemiology as a field of scientific inquiry largely began in the 19th century—although some examples of statistical analyses of public health data exist from before this time, including the work of John Graunt (1662). The first epidemiologic society, the *London Epidemiologic Society*, formed in 1850 (Morabia 2004), and John Snow published his foundational epidemiologic study, *On the Mode of Communication of Cholera,* in 1855 (Snow 1855). Since then, the field of epidemiology has rapidly expanded, both in size and scope. Today, epidemiology encompasses the study of all factors that influence the health of human populations (Porta 2014). While this includes the study of infectious diseases, many epidemiologists now study non-communicable diseases, environmental exposures, or social structures that lead to or increase disease and inequities. Some epidemiologists even study the impacts

■ *Eleanor J. Murray is Assistant Professor of Epidemiology, Boston University School of Public Health, Boston, Massachusetts. Her email address is ejmurray@bu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.4.105.

of economic policies, but the common focus of epidemiologists is on the goal of understanding and improving people's health (Krieger 2011).

Although epidemiology has evolved away from its roots in infectious diseases, epidemic response remains a core feature of epidemiologic expertise. However, in a changing and increasingly connected world, a successful response to COVID-19 is not entirely within the reach of even the best epidemiologic advice. As the pandemic continues to spread, the far-reaching consequences of COVID-19 are no longer just (nor perhaps even primarily) epidemiologic problems. After all, COVID-19 has impacted every sphere of life and aspect of society, and in many countries, decisions about how to mitigate the spread of COVID-19 have become highly politicized. Economists and others skilled in data analysis have recognized this and are anxious to put their skills to use, and those skills are desperately needed as the pandemic evolves.

As an epidemiologist, I ask economists interested in COVID-19 to build on their expertise and ours. Indeed, the efforts of economists in tackling the *economic* sequelae of this pandemic are vitally needed, as are the development of tools for tracking, predicting, and preventing future pandemics based on understanding the flow of people, goods, and other economic activity around the globe. But I also add that applying economics skills to evaluating epidemiology methods, as some more vocal economists have been doing, is not likely the best use of economic expertise nor will it be the most helpful for bringing COVID-19 under control. Despite having first encountered the concepts of infectious disease in January and February 2020, often through trial and error, many economists and other quantitative analysts have offered either prediction models of SARS-CoV-2 transmission or proposed strategies for reducing transmission. Many of these models failed to recognize the ways in which infectious diseases are very different from other types of quantitative data, and further, to recognize that expertise in modeling other datasets does not automatically translate to expertise in infectious disease modeling. As a result, many of the predictions or strategy proposals offered up by non-epidemiologists contain fundamental errors or oversights that greatly limit their value.

In this essay, I attempt to provide an epidemiologist's perspective on how economists can help improve our COVID-19 response. I begin with a discussion of how the goals of infectious disease modeling differ between applied and academic research settings and how some criticisms of epidemiology models are based on confusing these categories. I next discuss the tradeoff between data and assumptions in epidemiology. Early in a pandemic, an applied model must rely on a combination of limited data and assumptions. But as the pandemic evolves, it (perhaps counterintuitively) turns out that the quality of the data does not always improve and that key parameters may shift in unpredictable ways. The following section turns to some implications for epidemiology modeling and explains both what the public got wrong in interpreting these models and what epidemiologists got wrong in explaining them. I then identify some areas where work by economists could be especially helpful in the COVID-19 response.

Two years ago, this journal published, "An Economist's Guide to Climate Change Science," which offered a basic overview of the science written for economists (Hsiang and Kopp 2018). That article offers useful context and background for economists working on the benefits and costs of policies that seek to reduce greenhouse gas emissions, or seeking to estimate the economic consequences of rising temperatures and other climate effects. However, it would be a long road from that overview article to the frontier of atmospheric science models. Similarly, the companion paper by Avery et al. in this issue offers a clear and easy-to-follow basic introduction to the concepts of infectious disease epidemiology modeling for economists, but readers should not fall into the trap of assuming this gentle introduction, covering material that would be expected in the first few lectures of an introductory course on this topic, provides an exhaustive catalog of the methodological and substantive expertise required to conduct infectious disease modeling. Infectious disease epidemiology is not a new field. The principles of epidemic dynamics, prevention, and elimination are well-established and have been tested in disease outbreaks, large and small, as well as in computational models and laboratory experiments. There is no more reason for economists to jump into the production of epidemiology models than there is for them to become atmospheric scientists: after all, a central tenet of economics is the benefits that flow from specialization and exchange. I hope this discussion will provide a clearer understanding of the strengths and weaknesses of (infectious disease) epidemiology and spur more productive and impactful collaboration between our two disciplines.

Infectious Disease Epidemiology: Applied versus Academic

Should the public wear face masks? This question has been debated vigorously throughout the COVID-19 pandemic and has become increasingly politically charged. The answer to this question has also evolved over the course of the pandemic, with most public health officials now recommending face mask use for all individuals when outside the home. This changing and debated guidance has led to confusion among the general public and even to accusations that public health agencies and epidemiologists were concealing the benefits of face masks—although to what end they would do so remains unclear. The truth is that this question does not have a single answer, and the confusion arises from the distinction between two key facets of epidemiology.

Economics is often separated into theoretical and empirical work. In epidemiology, the concept of theory is used to discuss larger ideas of the interactions amongst the individual, environment, society, and biology of health and illness. What economists call "theory" would usually be referred to, instead, as "epidemiologic methods." Empirical epidemiology is thus intertwined in methods, based on the idea that we cannot understand why or how to conceptualize a model without understanding how the parameters arise. Rather than the theory/empirical distinction common in economics, a more natural distinction in epidemiology, especially

for infectious disease epidemiology, is between applied (or field) epidemiology and academic (or research) epidemiology.

Applied epidemiology focuses on providing rapid understanding to guide decision-making based on imperfect data, existing knowledge, and more than a little expert intuition (Rasmussen and Goodman 2019). Applied epidemiology is not restricted to infectious diseases, but infectious diseases often comprise a larger portion of the work of applied epidemiologists relative to the entire field of academic epidemiology. Even in academic epidemiology, epidemiologists recognize the trade-off between available data and required assumptions in disease modeling (Hernán 2015); this trade-off is even more apparent in applied epidemiology. During an outbreak, the limited available data must be offset with strong assumptions; the challenge is in understanding which assumptions are the most appropriate and which have the largest impact on model results and subsequent recommendations. The goal of applied epidemiology is to identify rapidly and under time pressure the category of infectious disease, to determine the key parameters specific to this infection, and to obtain data to update models based on those key parameters, while relying on prior experience, expertise, and intuition for selecting necessary assumptions. The bulk of COVID-19 models are applied epidemiology models. These models generally seek to estimate particular components of the prediction space, such as best- or worst-case scenarios.

Academic epidemiology, on the other hand, seeks to refine and hone a more detailed understanding of disease processes through extensive data collection, careful estimation of input parameters, and wide assessment of uncertainty (Murray et al. 2017; Eddy et al. 2012; Abuelezam, Rough, and Seage 2013). This process can take years and significant investment of human and financial resources. The goal of academic infectious disease modelers is to arrive at a more complete understanding of disease outbreaks that rely as much as possible on empirically obtained parameters rather than assumptions based on expert knowledge—although many disease models will require at least some input parameters that cannot ever be obtained directly from empirical data (Murray et al. 2020).

How does this apply to the changing guidance on community use of face masks? In January 2020, there was strong evidence supporting the use of personal protective equipment, including face masks, in high-risk settings such as health care facilities for the prevention of respiratory infections. However, the existing epidemiologic literature on the use of face masks by the general public for control of respiratory infections was extremely limited and showed mixed results (Brosseau 2020; Brosseau and Sietsema 2020; Chu et al. 2020). For example, one meta-analysis found that mask use in health care approximately halved the risk of influenza infection (Saunders-Hastings et al. 2017), and a randomized trial of non-pharmaceutical interventions in the home found an approximately 20 percent reduction in influenza infection for households using both face masks and hand sanitizer compared to hand sanitizer alone (Larson et al. 2010). In contrast, several randomized trials of households limited to face mask use alone had found no reduction in influenza transmission (Aiello et al. 2010; Canini et al. 2010; Cowling et al. 2008).

Lacking clear information on the benefits of community-level face mask use, epidemiologists in early 2020 engaged in internal discussion about the potential harms and benefits of this intervention, considering aspects such as the limited existing research, the limited supply and interrupted supply chains of masks, what was known at the time about the epidemiology of SARS-CoV-2 transmission, and concerns around the potential for "risk compensation" if people who were wearing masks then engaged in fewer other preventive measures (Bamber and Christmas 2020; Brosseau 2020; Brosseau and Sietsema 2020; Cheng 2020; Javid, Weekes, and Matheson 2020; King 2020). Based on these discussions, many applied epidemiologists, including those at the World Health Organization and Centers for Disease Control, initially advised against the use of face masks by the general public. Instead, they stressed the importance of hygiene and distancing-based interventions, such as hand-washing, social distancing, and quarantine.

Over time, however, new information emerged. First, it became clear that at least some subset of Americans would be amenable to wearing masks. Second, we learned that SARS-CoV-2 could be transmitted by individuals who were not (yet) symptomatic (Gandhi, Yokoe, and Havlir 2020). Finally, as the availability and use of both fabric and surgical masks increased, it became clear that even when individuals wearing masks did increase their risk behaviors (by, for example, joining protests), the evidence did not suggest that transmission in these settings was any higher than if attendees had been unmasked (Dave et al. 2020). Together, these observations have shifted most applied epidemiologists and public health officials towards encouraging the use of face masks by all individuals (Greenhalgh et al. 2020; Roderick et al. 2020).

However, this recommendation does not mean that the *academic epidemiology* of face mask usage by the general public during respiratory outbreaks has necessarily advanced much beyond what we knew in January 2020, and many academic epidemiologists remain agnostic about the value of face masks. In fact, if anything, it may be fair to say that academic epidemiologists have *fewer* answers about the science of face masks than we did 10 months ago—simply because we now have *more* questions. Previous research on face mask usage in respiratory outbreaks focused chiefly on evaluating either N95 masks or surgical masks, both of which are subject to regulatory standards. In contrast, many of the face masks used by the general public during the COVID-19 pandemic are made from fabric, both commercially and homemade, and the filtration efficacy of these masks is both unknown and potentially highly variable (Aydin et al. 2020; Davies et al. 2013; Tcharkhtchi et al. 2020). In addition, previous studies of face mask usage typically assumed individuals had been provided with training and guidance on how to appropriately don, doff, and wear face masks to maximize their benefits. In reality, adherence both in terms of frequency and correctness of face mask use is extremely variable among the general public. Despite this, existing attempts to model the population impacts of community-level face mask use have typically assumed perfect adherence and correct usage (Ferguson et al. 2020). Academic epidemiologists likely will be investigating and debating these topics for many years to come, both to fully characterize the

causal effect of community level mask-wearing strategies and to explore the actual risks and benefits that result from these (Bundgaard et al. 2020; Doung-ngern et al. 2020).

The distinction between applied and academic epidemiology arises from the different goals and time frames of these groups. Applied epidemiologists must provide the best available advice now and update as soon as new information is available—even if that advice or information arises more from experience and intuition than scientific fact. Academic epidemiologists, on the other hand, can afford to hold off making judgements until the best possible information is available and judgements can be supported by rigorous data collection and analyses. It is thus natural that these two groups will on occasion disagree about the best decision, as was the case with face masks, and such differences should not be misinterpreted as malicious or deceptive on the part of either group.

Data-Assumption Trade-Offs Change over Time as an Outbreak Progresses

Another challenging aspect of outbreaks that is well-known to epidemiologists, but may not be familiar to economists, is the complex and often counterintuitive ways in which data availability and quality can change over the course of an outbreak.

Initially, when a new disease emerges, very few data are available. As an outbreak progresses, the amount of data and the number of recognized cases naturally increase, but the completeness of the data does not necessarily increase in the same way. Many early attempts by non-epidemiologists (or epidemiologists with no experience in infectious diseases) to understand or predict COVID-19 went wrong when analysts either assumed that initial data would continue to describe the changes in disease spread over time, or that initial data could only be biased in one direction. In this section, I briefly describe some of the less intuitive ways in which data quality and biases can change over the course of an outbreak.

For diseases where mild cases or asymptomatic infections are rare, the earliest case reports are likely to be the most complete. This is because, although cases are easy to detect, the capacity for the system to track and record cases may degrade as the number of cases increases. Based on an analogy to the SARS outbreak of 2003, many initially thought this would be the case with COVID-19. This led to errors such as a focus on infection only among the elderly, tracking systems that recorded only death or recovery and failed to follow-up on longer term health outcomes, testing guidelines that restricted eligibility based on known contacts with COVID-19 patients, and diagnostic protocols which included only the symptoms most characteristic of early cases. Failure to update policies and guidance in these areas as the epidemiological and clinical picture of COVID-19 evolved may have contributed to the uncontrolled spread of SARS-CoV-2 in spring 2020.

For diseases with more common mild and asymptomatic cases, on the other hand, the completeness of the data can vary in complex ways over time. Early in

an outbreak, mild cases are likely to be missed, as happened with COVID-19, both because individuals are unaware they are ill and because mild cases are rare due to the small overall number of infections. As an outbreak increases, mild cases may continue to be undetected until a robust testing system has been implemented—again, this occurred with COVID-19 in many areas. As testing access increases, it can be challenging to understand how the prevalence of infection has changed since the mix of mild and severe cases detected will change. Without an understanding of the full scope and details of the case finding and reporting systems, infectious disease data can thus be extremely challenging to interpret. Again, this was observed during summer 2020 in the United States, where confirmed COVID cases began increasing post-lockdown due partly to an increase in testing availability detecting asymptomatic or mild cases among younger individuals. Although these outbreaks soon spread to older individuals who were more at risk of severe disease and death, the lag time between increasing test-detected COVID and increasing death rates led to intense speculation by the media that the overall severity of the disease had been misinterpreted, had changed, or had been overblown. Epidemiologists, on the other hand, were clear throughout the summer that the virus had not changed in any fundamental way.

Furthermore, it is not just that data change in their completeness over time—very often data *no longer exist* from which to estimate important epidemiologic parameters. This is particularly true for the basic reproductive number R_0. Early in an outbreak, the cases most likely to be identified are those severe enough to need hospitalization or medical attention, and mild or asymptomatic cases are largely overlooked, as are cases with atypical presentations. The messiness of this data means the R_0 estimate obtained from early data is very likely to be either an under- or overestimate of the true R_0. However, simply waiting until later in the pandemic is not necessarily a solution for obtaining an unbiased estimate of the R_0. In fact, it may not be possible to estimate R_0 from current data the further into a pandemic we get—the basic reproductive number specifically describes the number of new cases resulting from one infectious case introduced into an *entirely susceptible population*. Without a full accounting of asymptomatic, mild, and pre-symptomatic infections, our ability to identify an entirely susceptible population in which to estimate the R_0 can rapidly decay as a pandemic spreads.

For this very reason, while the reproductive number R_0 is an important tool for modeling outbreaks in an applied epidemiology setting, it is less commonly used to model infectious diseases in *academic* epidemiology research. Instead, given the benefit of time and the challenges of studying endemic diseases for which R_0 cannot be readily observed, academic epidemiologists often prefer to expend resources to measure contact rates and transmissibility per contact as well as how these vary by type of contact and by characteristics of individuals (as described by Avery et al. in their companion paper). The resulting academic models are often more complex in structure, frequently using agent-based or network-based transmission models.

Agent-based and network-based models are able to account for the full heterogeneity in transmission that occurs in the real world as well as the full spectrum of

characteristics which impact both exposure to infectious contacts and susceptibility to disease. They highlight areas of additional complexity which have long been recognized in epidemiology: for example, that the basic reproductive number, R_0, can be decomposed into the contact rate, the per-contact transmissibility rate, and duration of infectiousness; that the number of secondary cases per infection follows a distribution for which the mean may not always be the most appropriate summary measure; that heterogeneity in contact patterns, infectiousness, susceptibility, or other parameters may, if substantial, have a large impact on model results; and the need, in many cases, for finely stratified models.

Agent-based, network-based, and other complex system models can be specified in a broad range of ways that allow evaluation of the impacts of specific model components, assessment of interventions, or more realistic prediction of the evolution of an outbreak over time. However, while agent-based models, or other approaches incorporating this complexity, are sometimes used in applied epidemiology settings, the amount of time required to develop and validate these models can be prohibitive in an outbreak setting. Their use is therefore more common in academic epidemiology where these models can be carefully designed to help understand historic outbreaks and make predictions about future outbreak scenarios. Instead, applied epidemiologists often rely on compartmental models, which have limited capacity for incorporation of heterogeneity or endogeneity, but which can be more rapidly designed, tested, and applied to decision-making.

Clarifying the Goals of Early High-Profile Epidemiology Models: What the Public Got Wrong

COVID-19 models of early and mid-2020 are necessarily applied epidemiology models. These models have been rapidly developed based on the data at hand with the goal of providing insight into appropriate response and control activities. Despite this, many criticisms of these models seem to assume a goal closer to that of academic epidemiology—to create a detailed and highly accurate model of the full scope of the pandemic. Indeed, much of the public misunderstanding of epidemiology throughout this pandemic has involved a conflation of methods appropriate for applied epidemiology with those appropriate for academic epidemiology.

When critics argue over what high-profile epidemiology models "got wrong" about the COVID-19 pandemic, their analysis presupposes that the goal of these models was to predict, with both validity and accuracy, the actual total number of cases and deaths expected throughout the course of the pandemic under actual pandemic responses at both the individual and governmental levels. It is absolutely the case that both the high-profile Imperial College (Ferguson et al. 2020) and Institute for Health Metrics and Evaluation (IHME) models (Murray 2020) as well as all other current models, fell well short of this lofty goal; this is to be expected because it was not the intended goal of these models. The limitations of these models are well described in the Avery et al. essay in this issue, including the problems with model

structure, parameterization, and uncertainty. However, these criticisms reflect well-recognized limitations of mechanistic and phenomenological models.

But to put these concerns in real-world context, no infectious disease modeler expects to be able to accurately forecast the future based on sparse data from early in a pandemic. Even "nowcasting," the task of modeling the *current* number of true infections, is extremely challenging, especially early in a pandemic. Asking an infectious disease modeler to predict the exact trajectory of an outbreak is akin to asking an economist to select stocks for your portfolio or a climate scientist to predict the best day in 2022 for an outdoor wedding. These tasks, while of interest to many people, are not generally within the purview of scientists. Instead, the goal of both mechanistic and phenomenological models in epidemiology is to forecast a range of possible futures, given a specified set of assumptions.

In an outbreak setting, these early models help applied epidemiologists quickly evaluate the type of outbreak they are dealing with, narrowing the list of potentially appropriate actions to take and guiding the public health response. Academic epidemiologists, on the other hand, will likely spend many years attempting to create realistic models that explain exactly how and why the pandemic evolved the way it did; these models will then, in turn, be useful for future applied epidemiologists dealing with other pandemics.

In the case of the Imperial College model, two of the key assumptions which defined their original model were that the government would not respond to the COVID-19 pandemic with any interventions and that the general public would not respond to the pandemic with any changes to their own behavior. These assumptions are clearly unrealistic. However, by making these strong assumptions, the Imperial College model was able to provide epidemiologists and public health practitioners with a rapid estimate of the worst-case scenario: if SARS-CoV-2 was allowed to run unchecked through the population, what is the maximum amount of death that we might expect over the course of the outbreak until it burned out via herd immunity? The answer—510,000 deaths in the United Kingdom and 2,200,000 in the United States (Walker et al. 2020)—rightly spurred both governments and individuals to action.

In contrast, the IHME model used a phenomenological model with a different set of key assumptions to answer an entirely different question. In that model, restrictions on movement imposed by the US government were included so that the model could more accurately reflect the current case count. The goal of this model was to forecast as closely as possible, not *all* future cases, but *tomorrow's* cases—or rather, the expected number of hospital beds needed next week based on the number of cases expected to occur in the current week (Murray 2020). To achieve this goal as rapidly as possible, during the large initial surge in COVID-19 in the United States, the modelers made several simplifying assumptions: that the "lockdown" measures would continue unabated until the outbreak was completely eliminated from the United States and that the basic trajectory of cases over the past week was the best determinant of the trajectory of cases over the coming week. These assumptions are clearly simplistic and resulted in extremely unlikely longer-term forecasts,

including a prediction that elimination would happen by June 1, 2020. However, applied epidemiologists could use the short-term model forecasts from the IHME model to obtain a reasonable early warning on areas where hospital capacity was in danger of being overwhelmed, while ignoring the overly optimistic assumption of a June elimination.

The COVID-19 Infodemic: What Epidemiology Got Wrong

The COVID-19 pandemic is uncontained in the United States as of September 2020. As the scientific and applied field expected to protect the public from major health threats, epidemiology clearly has not succeeded in ending the pandemic. From the vantage point of six months since the US lockdowns began, it is clear that this is because many of the major challenges of the COVID-19 response were not, in fact, epidemiological. The science of epidemic response remains largely unchanged by COVID-19 from what it has been over the past century. Interventions such as hygiene, distancing, quarantine/isolation, and testing and contact tracing have worked in countries where they were systematically and rapidly deployed, such as New Zealand, Vietnam, and Mongolia. Instead, the major—and largely unforeseen—challenge for epidemiologists and public health professionals has been navigating the intense public scrutiny of the global scientific conversation about SARS-CoV-2/COVID-19 and the extreme politicization of that conversation.

Consider the conflict between public and scientific messaging about the American lockdowns in March and April. The main evidence-based approach to outbreak response advocated by epidemiologists at the time was, and remains, widespread frequent and rapid testing coupled with rigorous contact tracing, enforced quarantine and isolation, and appropriate personal protective equipment in all high-risk settings. However, delays in the availability of tests in the United States meant that targeted quarantine and isolation was unavailable as a response measure. Instead, the blunter tools of curfews and lockdowns were used to restrict transmission. Lockdowns were explained to the public as a tool for alternately eliminating transmission entirely or delaying transmission, resulting in confusion about the expected duration of both lockdown and the pandemic.

Many epidemiologists were vocal in the media and on social networks that the goal of lockdown should be to delay transmission until the availability of testing, contact tracing, quarantine supports, and personal protective equipment would allow safe reopening (Gottlieb et al. 2020). However, these more targeted approaches are still lacking in many jurisdictions, and the message much of the public seems to have internalized is that intervention to control the pandemic on a community-, state-, or federal-level is largely futile and that individual-level actions such as staying home and wearing a mask are sufficient (Gramlich 2020; Kramer 2020).

For a significant number of epidemiologists, myself included, COVID-19 represented their first experience communicating directly with the media and the

public. Public health communications are typically the purview of national and international organizations, such as the Centers for Disease Control and the World Health Organization, but these organizations failed to respond at the speed of the 24-hour news cycle. The infectious disease and applied epidemiology communities were focused on response, but gaps in expert communication remained which were rapidly filled by media personalities, talking heads, and non-epidemiologists.

At least from the perspective of this outsider, public communication appears to be a skill that economists have honed. Although surely many economists do not interface with the media or public regularly, news and public dialogue about economic topics frequently appears to include economists, and the public does appear to be aware of and defer to economic expertise. I suspect epidemiology has much to learn from economics about communicating with a skeptical and sometimes hostile public.

Epidemic-Related Economics, not *Epidemiology*-Related Economics

In the absence of more targeted control of COVID-19, the United States has relied largely on lockdown, resulting in unprecedented unemployment and a pandemic-induced recession. Epidemiologists as a discipline are singularly focused on saving lives but are generally unprepared to provide recommendations on how individuals can best *live* those lives. In particular, the pandemic-induced recession has raised many questions that are outside the scope of epidemiology and instead fall squarely in the domain of economics.

Macroeconomists as a group are already confronting the ways in which their models of fiscal policy, monetary policy, and financial regulation can be applied to a recession with very different underlying causes than, say, the Great Recession of 2007–2009 or the dot-com recession in 2001. The pandemic has reshaped concerns about the design of unemployment insurance, the connection between health insurance and employment, the availability of sick leave for workers, childcare, nursing home care, and many other issues. As an epidemiologist, I have been frustrated during this pandemic by the lack of answers to questions which lie beyond the usual boundary of epidemiology but are nonetheless vital for guiding pandemic response. Here, I highlight some examples of topics to which I believe economists could apply their unique skill sets, resources, and expertise to aiding in our understanding of and response to COVID-19.

1) Envisioning a Vibrant Remote Economy

Although official lockdowns are in the process of being lifted, much of the United States still functions in what might be called the "remote economy," where close proximity to large groups and to strangers are widely viewed as less desirable. Millions of people worldwide found themselves confined at home in March 2020, unable to continue to participate in economies which relied, in large part, on face-to-face contact. Indeed, even with lockdown (perhaps temporarily) eased or lifted

in many locations, many people are hesitant to resume pre-lockdown activities, and many businesses are realizing some potential financial benefits of remote workers. The consequences of the remote economy are potentially very widespread: in the workplace, in commuting patterns and the location of housing, in the economic health of cities, in the provision of medical advice and education, in the hospitality and leisure industries, and more. Indeed, the consequences of the remote economy may well outlast the COVID-19 pandemic. A full recovery of employment may require a dramatic shift across jobs and industries. A clear vision of this emerging remote economy is desperately needed.

2) Designing a Capitalism-Compatible Preparedness and Response Structure

Economists have long recognized that a market economy will tend to focus on activities that are likely to generate revenue. But being appropriately prepared to face a pandemic of uncertain form, whenever it arrives, is not this kind of activity. Similarly, responding to a pandemic cannot be sustainably done through market pressures alone. When SARS-CoV-2 began spreading outside Wuhan, China, it was predictable to many epidemiologists that supply chains of vital testing supplies, reagents, medications, and personal protective equipment would be overtaxed. Historically, public health agencies have sought to address this problem via the creation of pre-pandemic stockpiles. However, this system failed during COVID-19 when it was discovered that stockpiles created after the spread of epidemic H1N1 influenza in 2009–2010 had been left to expire or had their stock sold off. The COVID-19 epidemic was dramatically exacerbated by the lack of available testing and protective equipment. Epidemiologists do not, generally, have the skill set to untangle the complexities of so-called "just-in-time" supply chains and how they interact with regulatory imperatives. But economists might fruitfully consider how to build greater quick-response capacity into the economy.

3) Detecting Early Warning Signals of Pandemic Spread in Economic Networks

Epidemiologists interested in predicting pandemics have for many years tracked economic activity, especially trade in animals and animal products (Cunningham, Daszak, and Wood 2017). However, COVID-19 highlights that the costs and benefits of long-range international economic connections should come under greater scrutiny. For example, the Lombardy region of Italy includes one of the major world suppliers of nasopharyngeal swabs, Copan Diagnostics. In retrospect, it seems reasonable to expect that an outbreak of respiratory disease anywhere in the world might be accompanied by increased trade between Lombardy and the outbreak hotspots—and that even though this trade would be in the form of exports from Lombardy, it would involve increased contacts between individuals involved in the movement of supplies. Thus, it seems plausible that the Lombardy region of Italy could have been predicted as a location likely to be hit early and hard in any respiratory pandemic. Unfortunately, this is not the type of contact which epidemiologists are used to evaluating when predicting outbreak spread. Might economists be able to follow or even predict disease outbreaks by

looking at flows of trade in a way that would have allowed us to detect and intervene earlier to contain COVID-19 in Italy?

4) Collaborating to Evaluate Policy Impacts

While some epidemiologists are interested in the impacts of policy decisions on health, the bulk of epidemiology focuses on understanding the impacts of individual-level interventions. Economists, on the other hand, are much more experienced at understanding the impacts of group-level policy interventions. Economic methods for evaluating group-level interventions, such as difference-in-difference, interrupted time series, or instrumental variables, are less familiar to many epidemiologists. Economists could likely add much value through building partnerships and collaborations with epidemiologists who have subject-matter expertise that is vital to understanding potential endogenous variables and identifying key research questions. Such collaboration may not be easy—economists and epidemiologists have fundamental differences in language which can create confusion when trying to engage in cross-disciplinary collaborations. For example, epidemiologists rarely use the term "endogeneity," but understanding, assessing, and controlling for endogenous variation is at the heart of nearly all epidemiologic investigations—often under terms like "confounding" and "selection bias."

5) Rapidly Deploying Research Resources

Finally, economists have access to research resources which could be rapidly deployed to answer important questions about how, where, and when people are experiencing effects of a pandemic, beyond case counts and deaths. For example, economic research can be useful in tracking the different ways in which people react to the pandemic, thus helping policymakers to identify groups where the burden of the pandemic may be especially large.

As one example, Alsan et al. (2020) surveyed US adults to understand how knowledge, behaviors, and incidence of COVID-19 were distributed, with particular attention to the ways in which existing social, economic, and health disparities might be exacerbated during the pandemic. Their results, although perhaps unsurprising to many in public health, provided important quantitative evidence of COVID-19 related disparities that can help guide targeted messaging, outreach, and intervention. As another example, the United States has relied on so-called "essential workers" during the pandemic, which puts such workers at higher risk for contracting SARS-CoV-2. In addition, many of the occupations deemed essential are blue collar occupations. McCormack et al. (2020) confirmed the suspicion of many epidemiologists and public health practitioners that essential workers were often economically vulnerable, lived in smaller and more crowded living conditions, and belonged to minority groups. This research will serve an important role in helping guide public health actions to protect the economic health of these vulnerable populations. Of course, a number of economists are already working in these areas, and I look forward to seeing more of this work in the future.

Conclusion

Epidemiology is a sister discipline of economics; both fields combine the study of observational and experimental data on human populations with a goal of understanding and supporting human flourishing. Both our fields are forced to grapple with observational data in ways that other scientific fields can more easily sidestep by experimentation, and both our fields have developed robust approaches to combining this data with subject-matter expertise to make causal inferences. Both our fields require a deep understanding of sociological processes and quantitative methods. Both fields can even trace the origin of key methods to John Snow—the originator of the difference-in-difference approach and a founder of epidemiology.

Despite this, economics and epidemiology have largely operated in isolation. We have developed our own, often mutually unintelligible, languages and practices, and we have allowed the public to create a narrative of conflict between us (Escandón et al. 2020). But we need not choose between a healthy public and a healthy economy! In the face of a systemic catastrophe like COVID-19, we can no longer afford to work alone. We need to join forces to recognize and apply our discipline-specific strengths to the problem at hand.

■ *The author would like to thank Kareem Carr for his thoughtful comments on an earlier version of this manuscript as well as the editors at the Journal of Economic Perspectives for providing the opportunity to write this commentary.*

References

Abuelezam, Nadia N., Kathryn Rough, and George R. Seage III. 2013. "Individual-Based Simulation Models of HIV Transmission: Reporting Quality and Recommendations." *PLoS One* 8 (9).

Aiello, Allison E., Genevra F. Murray, Vanessa Perez, Rebecca M. Coulborn, Brian M. Davis, Monica Uddin, David K. Shay, Stephen H. Waterman, and Arnold S. Monto. 2010. "Mask Use, Hand Hygiene, and Seasonal Influenza-Like Illness among Young Adults: A Randomized Intervention Trial." *The Journal of Infectious Diseases* 201 (4):491–98.

Alsan, Marcella, Stefanie Stantcheva, David Yang D, and David Cutler. 2020. "Disparities in Coronavirus 2019 Reported Incidence, Knowledge, and Behavior among US Adults." *JAMA Network Open* 3 (6).

Aydin Onur, Bashar Emon, Shyuan Cheng, Liu Hong, Leonard P. Chamorro, and M. Taher A. Saif. 2020. "Performance of Fabrics for Home-made Masks against the Spread of COVID-19 through Droplets: A Quantitative Mechanistic Study." *Extreme Mechanics Letters* 40.

Bamber, Hames H., and Tracey Christmas. 2020. "Covid-19: Each Discarded Face Mask Is a Potential Biohazard." *BMJ*.

Brosseau, Lisa. 2020. "COMMENTARY: COVID-19 Transmission Messages Should Hinge on Science." *CIDRAP*, March 16. https://www.cidrap.umn.edu/news-perspective/2020/03/commentary-covid-19-transmission-messages-should-hinge-science.

Brosseau, Lisa M., and Margaret Sietsema M. 2020. "COMMENTARY: Masks-for-All for COVID-19 Not Based on Sound Data." *CIDRAP*, April 1. https://www.cidrap.umn.edu/news-perspective/2020/04/commentary-masks-all-covid-19-not-based-sound-data.

Bundgaard, Henning, Johan Skov Bundgaard, Daniel Emil Tadeusz Raaschou-Pedersen, Anton Friis Mariager, Natasja Schytte, Christian von Buchwald, Tobias Todsen, et al. 2020. "Face Masks for the Prevention of COVID-19 - Rationale and Design of the Randomised Controlled Trial DANMASK-19." *Danish Medical Journal* 67 (9): 1–10.

Canini Laetitia, Laurent Andréoletti L, Pascal Ferrari, Romina D'Angelo, Thierry Blanchon, Magali Lemaitre, Laurent Filleul, et al. 2010. "Surgical Mask to Prevent Influenza Transmission in Households: A Cluster Randomized Trial." *PLoS ONE* 5 (11).

Cheng, Sheung-Tak. 2020. "Covid-19: Are Face Masks a Good Long Term Strategy?" *BMJ* 369.

Chu, Derek K., Elie A. Akl, Stephanie Duda, Karla Solo, Sally Yaacoub S, and Holger J. Schünemann. 2020. "Physical Distancing, Face Masks, and Eye Protection to Prevent Person-to-Person Transmission of SARS-CoV-2 and COVID-19: A Systematic Review and Meta-Analysis. *The Lancet* 395 (10242): 1973–87.

Cowling, Benjamin J., Rita O.P. Fung, Calvin K.Y. Cheng, Vicky J. Fang, Kwok Hung Chan, Wing Hong Seto, Raymond Yung, et al. 2008. "Preliminary Findings of a Randomized Trial of Non-Pharmaceutical Interventions to Prevent Influenza Transmission in Households." *PLoS ONE* 3 (5).

Cunningham, Andrew A., Peter Daszak, and James L.N. Wood. 2017. "One Health, Emerging Infectious Diseases and Wildlife: Two Decades of Progress?" *Philosophical Transactions of the Royal Society B Biological Sciences* 372 (1725).

Dave, Dhaval M., Andrew I. Friedson, Kyutaro Matsuzawa, Joseph J. Sabia, and Samuel Safford. 2020. "Black Lives Matter Protests, Social Distancing, and COVID-19." NBER Working Paper 27408.

Davies, Anna, Katy-Anne Thompson, Karthika Giri, George Kafatos, Jimmy Walker, Allan Bennett. 2013. "Testing the Efficacy of Homemade Masks: Would They Protect in an Influenza Pandemic?" *Disaster Medicine Public Health Preparedness* 7 (4): 413–18.

Doung-ngern, Pawinee, Repeepong Supanchaimat, Apinya Panjangampatthana, Chawisar Janekrongtham, Duangrat Ruampoom, Nawaporn Daochaeng, Napatchakorn Eungkanit, et al. 2020. "Case-Control Study of Use of Personal Protective Measures and Risk for Severe Acute Respiratory Syndrome Coronavirus 2 Infection, Thailand." Emerging Infectious Diseases 26 (11).

Eddy, David M., William Hollingworth, J. Jaime Caro, Joel Tsevat, Kathryn M. McDonald, and John B. Wong. 2012. "Model Transparency and Validation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7." *Value in Health* 15 (6):843–50.

Escandón, Kevin, Angela L. Rasmussen, Isaac I. Bogoch, Eleanor Murray, Karina Escandón. 2020. "COVID-19 and False Dichotomies: Time to Change the Black-or-White Messaging about Health, Economy, SARS-CoV-2 Transmission, and Masks." *OSF Preprints* https://osf.io/k2d84/.

Ferguson, N., D. Laydon, G.V. Gilani, Natsuko Imai, Kylie E.C. Ainslie, M. Baguelin, A. Boonyasiri, et al. 2020. "Report 9: Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID19 Mortality and Healthcare Demand." *Imperial College London* https://doi.org/10.25561/77482.

Gandhi, Monica, Deborah S. Yokoe, Diane V. Havlir. 2020. "Asymptomatic Transmission, the Achilles' Heel of Current Strategies to Control Covid-19." *New England Journal of Medicine* https://www.nejm.org/doi/full/10.1056/NEJMe2009758.

Gottlieb, Scott, Caitlin Rivers, Mark B. McClellan, Lauren Silvis, and Crystal Watson. 2020. "National Coronavirus Response: A Road Map to Reopening." *American Enterprise Institute Papers & Studies.*

Gramlich, John. 2020. Why Are COVID-19 Cases Rising in U.S.? Republicans Point to More Testing, Democrats to More Infections. *Pew Research Center*, August 14. https://www.pewresearch.org/fact-tank/2020/08/14/why-are-covid-19-cases-rising-in-u-s-republicans-point-to-more-testing-democrats-to-more-infections/.

Graunt, John. 1662. *Natural and Political Observations Made upon the Bills of Mortality.* London: Royal Society of London.

Greenhalgh, Trisha, Manuel B. Schmid, Thomas Czypionka, Dirk Bassler, Laurence Gruer. 2020. "Face Masks for the Public during the COVID-19 Crisis." *BMJ* https://doi.org/10.1136/bmj.m1435.

Hernán, Miguel A. 2015. "Invited Commentary: Agent-Based Models for Causal Inference—Reweighting Data and Theory in Epidemiology." *American Journal of Epidemiology* 181 (2): 103–05.

Hsiang, Solomon, and Robert E. Kopp. 2018. "An Economist's Guide to Climate Change Science." *Journal of Economic Perspectives* 32 (4): 3–32

Javid, Babak, Michael P. Weekes, and Nicholas J. Matheson. 2020. "Covid-19: Should the Public Wear

Face Masks? DOI: 10.1136/bmj.m1442 .

King, Frances M. 2020. "Covid-19: Face Masks Could Foster Distrust and Blame." https://doi.org/10.1136/bmj.m2009.

Kramer, Stephanie. 2020. "More Americans Say They Are Regularly Wearing Masks in Stores and Other Businesses." *Pew Research Center*, August 27. https://www.pewresearch.org/fact-tank/2020/08/27/more-americans-say-they-are-regularly-wearing-masks-in-stores-and-other-businesses/.

Krieger, Nancy. 2011. *Epidemiology and the People's Health.* Oxford: Oxford University Press.

Larson, Elaine L., Yu-Hui Ferng, Jennifer Wong-McLoughlin, Shuang Wang, Michael Haber, Stephen S. Morse. 2010. "Impact of Non-Pharmaceutical Interventions on URIs and Influenza in Crowded, Urban Households." *Public Health Reports* 125 (2):178–9.

McCormack, Grace, Christopher Avery, Ariella Kahn-Lang Spitzer, and Amitabh Chandra. 2020. "Economic Vulnerability of Households with Essential Workers." *Jama* 324 (4): 388–90.

Morabia, Alfredo, ed. 2004. *A History of Epidemiologic Methods and Concepts.* Basel, Switzerland: Birkhäuser Verlag.

Murray, Christopher J.L. 2020. "Forecasting COVID-19 Impact on Hospital Bed-Days, ICU-Days, Ventilator-Days and Deaths by US State in the Next 4 Months." *medRxiv* https://doi.org/10.1101/2020.03.27.20043752.

Murray, Eleanor J., James M. Robins, George R. Seage III, Kenneth A. Freedberg, Miguel A. Hernán. 2017. "A Comparison of Agent-Based Models and the Parametric G-Formula for Causal Inference." *American Journal of Epidemiology.* 186 (2): 131–42.

Murray, Eleanor J., James M. Robins, George R. Seage III, Kenneth A. Freedberg, Miguel A. Hernán. 2020. "The Challenges of Parameterizing Direct Effects in Individual-Level Simulation Models." *Medical Decision Making* 40 (1): 106–11.

Porta, Miguel, ed. 2014. *A Dictionary of Epidemiology.* Oxford: Oxford University Press.

Rasmussen, Sonja A., and Richard A. Goodman, eds. 2019. *The CDC Field Epidemiology Manual.* Oxford: Oxford University Press.

Roderick, Paul, Guiqing Lily Yao, Nisreen A. Alwan, Iain Buchan, Rochelle A. Burgess, Tim Colbourn, Anthony Costello, et al. 2020. "Open Letter Recommending Use of Cloth Face Coverings in Public Places." https://docs.google.com/forms/d/e/1FAIpQLSeS_4rQIsqNYonGF2VjazPvCzrlIB6NtJ_N-8ybqqOspaqAAg/viewform. (accessed September 18, 2020).

Saunders-Hastings, Patrick, James A.G. Crispo, Lindsey Sikora, and Daniel Krewski. 2017. "Effectiveness of Personal Protective Measures in Reducing Pandemic Influenza Transmission: A Systematic Review and Meta-Analysis." *Epidemics* 20: 1–20.

Snow, John. 1855. *On the Mode of Communication of Cholera.* London: John Churchill.

Tcharkhtchi, A., N. Abbasnezhad, M. Zarbini Seydani, N. Zirak, S. Farzaneh, M. Shirinbayan. 2020. "An Overview of Filtration Efficiency through the Masks: Mechanisms of the Aerosols Penetration." *Bioactive Materials* 6 (1):106–22.

Walker, Patrick G.T., Charles Whittaker, Oliver Watson, Marc Baguelin, Kylie E.C. Ainslie, Sangeeta Bhatia, Samir Bhatt, et al. 2020. *Report 12: The Global Impact of COVID-19 and Strategies for Mitigation and Suppression* London: Imperial College COVID-19 Response Team.

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 121–145

A 30-Year Perspective on Property Derivatives: What Can Be Done to Tame Property Price Risk?

Frank J. Fabozzi, Robert J. Shiller, and Radu S. Tunaru

Movements in property prices can pose severe risks to those who hold real estate assets as well as to the financial sector and even macroeconomic stability. A vivid example is the global financial crisis of 2007–2009, when approximately eight million American homes were foreclosed, and $7 trillion dollars in home equity vanished (in this journal, Goodman and Mayer 2018). The sharp decline in US housing prices and how it echoed through the US financial system was a primary driver of the Great Recession of 2007–2009 (Gertler and Gilchrist 2018). This connection goes well beyond the US economy. Many past financial crises have shown a connection to house-price risk because irrational and exuberant periods are often paired with property booms and bubbles. It is well known that in many major economies, house-price growth is related to financial stability, particularly in those countries that use variable-rate mortgages and market-based property valuation for mortgage loans (Tsatsaronis and Zhu 2004). The interaction of housing price, household debt, and the financial sector can help explain how the Great Recession affected high-income countries around the world as well as explain economic fluctuations around the world going back to the 1970s (for example, see Mian and Sufi 2018).

Yet the financial instruments available in financial markets to control this globally omnipresent risk remain in a state of infancy. For example, consider a

■ *Frank J. Fabozzi is a Professor of Finance, EDHEC Business School, Nice, France. Robert J. Shiller is Sterling Professor of Economics, Yale University, New Haven, Connecticut. Radu Tunaru is Professor of Finance and Risk Management, University of Sussex Business School, Brighton, United Kingdom. Their email addresses are frank.fabozzi@edhec.edu, robert.shiller@yale.edu, and r.tunaru@sussex.ac.uk.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.4.121.

homebuyer circa 2005 or 2006 who faced conflicting expert opinions about whether housing prices might fall in the near-term. For example, McCarthy and Peach (2004) and Himmelberg, Mayer, and Sinai (2005) argue that house prices were not going to collapse. Conversely, Shiller (2006) presents ample historical evidence that house prices at the time were far from the norm suggested by historical patterns, and Shiller (2005) writes: "Significant further rises in these [housing] markets could lead, eventually, to even more significant declines." But that concerned homebuyer had no mechanism to hedge against the risk that the price of the specific house they were purchasing would decline. Nor did investors purchasing mortgage-backed securities have any straightforward way to hedge against the risk of a widespread decline in the average property prices, either within an urban area, region, or nationally.

The main objective of this paper is to offer a perspective on the principal obstacles hindering the development of financial derivatives based on real estate prices—especially housing prices—and what could be done to overcome these difficulties. By the late 1980s, Case and Shiller (1989) started a research agenda dedicated to the search for financial solutions that could mitigate house price risk.

We first provide an overview of some basic financial derivatives and their benefits. We then discuss some history about the volatility of property prices and their interaction with the financial sector. Between 1870 and the middle of the 20th century, available data suggest that real house prices remained fairly stable in many places worldwide (Jordà, Schularick, and Taylor 2015). Some financial derivatives were then developed to reduce the risk for originators of mortgages. But the lack of development of financial derivatives based directly on property prices probably had only a modest effect up to the 1970s, when a combination of inflationary and real increases in housing prices shook up these markets. We discuss early attempts to create property derivatives in the 1990s, which either failed or were very limited in scope. However, after 2006 a functional market did emerge in real estate derivatives, both in the United States (on housing) and in the United Kingdom (on commercial property). We discuss the empirical evidence on benefits that have arisen from these financial derivatives as well as their flaws and limitations.

Finally, we then discuss the main specific obstacles to a more complete development of a property derivatives market: problems in matching a suitable property index to the property derivatives themselves, concerns about a limited number of parties in the market, problems of modelling property derivatives, and concerns about how regulations may affect the participation of large financial institutions in these markets. Our study is complementary to the review of housing finance in Shiller (2014), where the slow pace of innovation with respect to the development of tools for controlling property risk is also criticized.

Advantages of Real Estate Derivatives

A number of studies starting with Case and Shiller (1989) have pointed out the benefits of introducing property derivatives—for housing prices in particular.

Table 1
Some Examples of Simple Financial Derivatives

Derivative	*Definition*	*Example*
Futures and forwards	An agreement to buy or sell a certain asset at a certain price at a specific date in the future.	An airline, agreeing to buy fuel in the future at a certain price; a farmer, agreeing to sell a crop in the future at a certain price. If the contract is on an exchange there are margin payments subject to a marking-to-market process that must be paid in order to remove counterparty risk, and this is a futures contract. If the contract is directly between two parties over-the-counter then the contract is paid off only at maturity. Futures are standardized while forwards are bespoken.
Call option	The right to purchase an underlying security at a predetermined future time and "strike price."	Some companies grant workers a call option to purchase company stock in the future at a certain strike price.
Put option	The right to sell an underlying security at a specified price and date.	A put option on the S&P 500 can be exercised by the holder if the value of the S&P500 at some future maturity falls below the chosen exercise threshold. Hence, an asset manager may recoup portfolio losses by buying a put as a hedge against a market decline.
Total return swap	The "payer" agrees to send the total return from a certain asset to the "receiver," and the "receiver" agrees to either a make a payment based on a benchmark interest rate or the total return on a different asset.	An asset manager agrees to pay a funding market floating rate plus a fixed spread to a counterparty in return for the return on some property index. Basically, the asset manager creates a leveraged position in the property index funded at LIBOR plus the fixed spread.

Examples focused on the US market include Shiller (1993c) and Shiller and Weiss (1999), while Case, Shiller, and Weiss (1991, 1993) explain the need for house price index futures and options. For the case of the United Kingdom, Gemmill (1990) argues for the benefits of futures trading for the house price market in the United Kingdom and Baum (1991) make the case for commercial property futures (see also Thomas 1996). Table 1 provides a quick definition of some common financial derivatives. In this section, we discuss four types of gains that financial derivatives provide: 1) improved information about the path of future prices, 2) hedging against risk, 3) a tool for broadening investment portfolios, and 4) a basis for new financial products.

Information about the Future of Evolution of Property Prices

The principal role of property derivatives is that they allow end-users to extract more reliable information about the future evolution of property prices. Of course, for the introduction of property derivatives (residential and commercial) to be successful, their usage must appeal to sophisticated market players who find it advantageous to take on property risk such as hedge funds, pension funds,

insurance funds and speculators. Such market players provide deep pockets that help to keep the market liquid along with deep expertise in valuation of assets.

Voicu and Seiler (2013) and Uluc (2018) investigate the theoretical impact of introducing house price futures on managing house price risk. For example, Uluc (2018) adapts the De Long et al. (1990) model of noise traders to the housing market: that is, the market is made up of noise traders who have imperfectly predictable beliefs about waves of property values up and down and sophisticated investors who try to predict the noise traders—and in doing so, the sophisticated investors can magnify the size of the waves. The model demonstrates that there are three channels by which house price futures may affect house prices. First, the noise traders who are looking to benefit from momentum in prices no longer need to purchase houses themselves but can now focus on trading the financial derivative of housing futures. The market for buying and selling actual houses is slow and somewhat illiquid, while trades in the synthetic (paper) domain are settled very quickly. Depending on the noise traders' perception of the market, it is possible for house price futures trading to trigger either an increase or decrease in the price volatility of residential property. Second, a housing futures market allows for short selling. Thus, when noise traders begin to display irrational exuberance, they become the only market players buying in expectation of higher housing prices, while sophisticated (knowledgeable) households and investors use house price futures for short-selling for the investment part of the housing asset—and during that time are more likely to rent than to buy. Third, Uluc (2018) argues that when house price futures overall become attractive to sophisticated investors, the volatility of house prices decreases. Moreover, the presence of sophisticated risk-neutral investors in this model will in the long run eliminate the imperfections and distortions in the housing market.

A market in property derivatives can also clarify certain prices that are now bundled together. For example, Case and Shiller (1996) explain how financial institutions and investors could use futures and options to extract information and manage two major interrelated risks that lenders face: price risk and default risk.

Hedging Housing Price Risk

Property derivatives offer certain end users the opportunity to hedge or control property-related risks—perhaps especially those concerned about a price fall. An obvious use would be that some homebuyers, at the time they purchase their house, might wish to purchase an insurance policy against the risk of a decline in house prices. Shiller and Weiss (1999) expanded this idea to explain how residential property futures would facilitate the selling of home equity insurance. Hence, house price futures, for example, could help insurance companies remove the risk of endemic house price declines. The existence of house price futures may allow the coexistence of another financial innovation: insurance that you will not lose the value of the down payment on your home and even price guarantees on new homes. Shiller (1993c) put forth theoretical arguments for this financial innovation which almost ten years later proved to be effective in 2002 in Syracuse, New York, when a home equity insurance program was launched (Caplin et al. 2003).

In addition, residential property derivatives could be employed by mortgage insurers to hedge the risk of higher defaults that occur when housing prices fall (Case et al. 1993; Case and Shiller 1996). All major banks are now required to pass regular stress tests imposed by regulators and some of these tests involve severe house price market collapses. Thus, using house price derivatives could be a solution to mitigate the banks' overall portfolio downside risk exposure in order to pass stress tests.

One can also imagine those who wish to hedge against a rising price of housing. Say that a young household lives in an area where real estate prices seem to be rising faster than incomes, but they are not yet ready to purchase a home. By purchasing housing futures based on an index of local real estate prices, or an option, they can reduce the chance of being "priced out" of a real estate market where prices are rising.

Portfolio and Investment Decisions

When seeking to optimize an investment strategy, property derivatives provide an additional tool with distinctive characteristics of risk and return. For example, some investors might seek to acquire exposure to real estate profit and losses in a synthetic manner, in this way obtaining exposure to real estate sectors where it would be almost impossible to trade on the spot market (like shopping centers and warehouses). Other possible actions could involve moving between asset classes or sectors or "relative value trading" where the investor seeks to benefit from a change in the spread between the outcome of the property derivative and some other asset. As Englund (2010) pointed out, property derivatives would enable households to disentangle their housing consumption decisions from their housing investment decisions: for example, a renter could use property derivatives to benefit from rising real estate prices.

The development of property derivatives markets has been hampered by the fact that, for long periods of time, property markets were one-sided—specifically, with a lack of investors willing to be counterparties in property derivatives transactions. But with a market for property derivatives, investment banks and investment funds should be willing to buy the property risk because they cannot plausibly claim that they are fully diversified without holding positions in property markets. For investors who want to use property derivatives for managing their exposure to this asset class, one of the major problems in trading in this asset class is the lack of fungibility and the implied impossibility of short selling of the spot asset. As noted earlier, not being able to short-sell an asset when market values appear to be inflated relative to fundamentals may be a direct contributor to increased market sentiment, ultimately resulting in real estate bubbles.

New Financial Products: Reverse Mortgages

The discussion has already mentioned how property derivatives can provide a basis for new financial products like down-payment insurance. Another emerging financial product where housing price risk is highly relevant is "reverse mortgages,"

in which a homeowner receives periodic payments for a fixed period or life, secured by the value of the property that will be sold after death. Reverse mortgages may be especially beneficial for elderly households with low-income, poor health, and limited non-housing wealth (Nakajima and Telyukova 2017). The UK equivalent of a reverse mortgage is called an "equity release mortgage." It allows the borrower, a senior person over a certain age, to get a lump-sum or to draw regular or when-needed sums from a credit line. The loan accrues interest that will be paid only at the termination event, when the borrower dies, moves into long-term care, or prepays. The loan can be granted to individuals or couples who own and live in a house that is used as a collateral to pay back the loan at termination.

However, an obvious concern for the reverse mortgage market is the risk of a decline in property values. In a US-style reverse mortgage, the issuer must embed in the deal an insurance policy against a house price decline. This insurance policy serves several functions: that the borrower will not absorb the negative equity when the loan is terminated, that the loan will continue to pay its installments even if the lender goes bankrupt, and that the reinsurer will pay the lender if the negative equity guarantee insurance policy is triggered. In essence, these products depend crucially on a put option on the house price at an uncertain time in the future (which can be priced depending on the actuarial characteristics of the borrower), along with a strike price equal to the accumulated balance at a fixed rate. In the United Kingdom, the regulator requires the insurer issuing the loan to cover the risk that the house price at termination is lower than the loan balance to be repaid.

Because an expanding market for reverse mortgages is highly dependent on the no-negative equity guarantee, house price hedging instruments would help to satisfy regulators and help improve the linkages between the housing market and the health care market (Tunaru 2017). This problem could be solved immediately if a futures contract on house prices existed, along with the use of put options.[1] But with such markets still at their early stages, the price of providing a no-negative equity guarantee remains higher than it needs to be, thus impeding growth of the market for reverse mortgages.

Why Did the Need for Real Estate Derivatives Rise in the 1970s?

Over time, many assets have served as a basis for derivative contracts, including stock prices, bond prices, and commodities prices like oil or wheat. One reason why house prices have not done so is that house prices in the majority of developed

[1] In the UK market for equity release mortgages, for example, regulators accept the application of the Black-Scholes (1973) option pricing model to evaluate the risks. However, it is well-accepted among researchers and industry that with this approach, the valuations of the necessary house price options are quite inflated due to the way the Black-Scholes model builds in volatility. Regulators naturally prefer to be conservative. In contrast, insurers argue that the very high capital limits imposed are impeding the development of this market. Issues with appropriate modelling of property derivatives will be discussed in more detail later in this paper.

economies remained relatively constant in real terms from the 19th century to the 1960s (Knoll, Schularick, and Steger 2017). Of course, there's a lot of detail buried in that word "relatively." Knoll, Schularick, and Steger (2017) provide ample evidence that house prices stayed stable before World War I, although income per capita increased and then, relative to income, they decreased until the 1960s. Glaeser (2013) argues that the United States has been, for a long time, a nation of property speculators, with local and regional boom-bust periods that created substantial social costs and financial instability. However, these movements in housing prices often evened out in the long run. As one example, using a repeat sales index based on 86 properties in New York City's borough of Manhattan over a century, Wheaton, Baranski, and Templeton (2009) show that in every decade, property prices increased between 20 and 50 percent and then declined the same way such that in the late 2000s, real estate in that city was worth almost the same as at the turn of the 19th century in real terms. The international historical evidence suggests that, since 1870, house prices in Australia outpaced income; in the United States and European countries like Belgium, Sweden, and Germany, house price growth was substantially behind income growth; and for Canada, Japan, and the United Kingdom, house prices grew more or less in line with income (Jordà, Schularick, and Taylor 2015).

Still, the fact that house prices in high-income countries were "relatively" constant for a sustained time means that the perceived need for financial derivatives to hedge against movements in these prices was not large. This may explain why there was little motivation for introducing futures contracts related to house prices.

Moreover, there is a long history of other financial instruments that protected the originators of mortgage loans from risk of fluctuating housing prices by creating financial securities based on a pool of mortgage loans, which can then be sold to investors. The primary historical example is covered bonds: these debt securities are specialized instruments issued by financial institutions under specific legislative measures. Covered bonds are basically a hybrid between corporate bonds and mortgage-backed securities. The collateral for a covered bond is a pool of mortgage loans (commercial and residential) and some other public sector assets. The payments to the holders of covered bonds are a liability of the issuer. Covered bonds may receive a credit rating higher than the issuer's credit rating, although the credit rating mainly reflects the issuer. This apparent paradox can be explained by the legislative measures supporting covered bonds. Unlike corporate bonds, the mortgage loans are segregated to the benefit of the security holders so that the credit of the covered bonds also depends on the credit quality of the collateral. Covered bonds were first issued in Germany (then Prussia) in 1769 where they are called *pfandbriefe*, and in 1797 in Denmark where they are called *realkreditobligationer* (Kothari 2012). Covered bonds are commonly used today in many European countries. They are now issued in Australia, New Zealand, and Canada, but rarely issued in the United States (although after the subprime mortgage crisis, the US Department of the Treasury formulated a plan to promote the issuance of covered bonds). From a global perspective, covered bonds constitute the largest bond market after the US bond market.

Table 2
Two Forms of Property Loans Securitization: Covered Bonds and Mortgage-Backed Securities

Characteristic	*Covered bonds*	*Mortgage-backed securities*
Asset composition	Defined by law and substitutable	Cannot be exchanged after issuance
Support from issuer	Allowed use of other assets by issuer	Not allowed
Balance sheet	On issuer's balance sheet	Off balance sheet
Issuer's absorption of default risk	Yes	Only pro-rata to their equity tranche
Issuer's absorption of prepayment risk	Yes	No
Number of issuances from one collateral pool	Multiple issuances allowed	One pool one issuance.

In Table 2, we contrast the two main forms of property loans securitization, covered bonds and the mortgage-backed securities more common in the United States. In a US context, mortgage-backed securities have a much shorter history. US mortgage securitization was first used in the 1920s by insurance companies who issued mortgage participation certificates backed by a pool of mortgage loans that they guaranteed (Goetzmann and Newman 2010). Up until the real estate crash that accompanied the Great Depression, these securities were actively traded.

The Great Depression triggered spectacular innovations in mortgage designs in the United States. Until that time, mortgages were not fully amortized; instead, they were balloon instruments in which the principal was only partially amortized (or not amortized at all) at maturity. Thus, the end of the mortgage left the homeowner with the problem of refinancing the balance and exposed the lender to considerable default risk. Sometimes the lender (typically a depository institution at the time) had the power to require repayment of the outstanding balance on demand or upon relatively short notice, even if the mortgager had been making payments on time. This type of mortgage designed proved disastrous during the Great Depression and contributed to both its depth and personal distress, as banks afflicted by losses on their loans and by depositors' withdrawals found it necessary to liquidate their mortgage loans at a time when debtors found it impossible to refinance. The disastrous experience led to the widespread adoption of the current fixed-rate, level payment, fully amortized mortgage by the mid-1930s. As Fabozzi and Modigliani (1992) note, the level-payment mortgage was a great success, contributing to the recovery of the housing market after the Great Depression. This mortgage design continued to perform a valuable role in financing residential real estate in the first two decades of the post-World War II period until the inception of the era of high inflation in the 1970s.

During this time, one of the main changes in housing finance was an effort to develop a secondary mortgage market in the United States. The Government National Mortgage Association (Ginnie Mae) and the Federal National Mortgage Association (Fannie Mae) were created in 1968 and then the Federal Home Loan Mortgage Corporation (Freddie Mac) followed in 1970. These three entities worked with qualified mortgage originators to create mortgage-backed securities that were guaranteed by one of these entities. These "agency mortgage pass-through securities" represented the first generation of mortgage-backed securities. Later, other mortgage originators issued securities that were backed solely by the credit quality of the underlying mortgage pool, referred to as private-label, mortgage-backed securities. It was not until the 1990s that the first mortgage-backed securities backed by a pool of commercial mortgage loans were issued.

However, the fixed-rate, level payment, fully amortized mortgage—and the mortgage-backed securities based on it—were unprepared when the inflation of the 1970s produced devastating effects on the housing industry in all countries.[2] Adjustable-rate mortgages shifted the risks of inflation to borrowers, rather than lenders. But the rise in mortgage interest rates that accompanied, and roughly matched, the rise in inflation pushed homeownership out of the reach of major segments of the population—notably the young and the first-time homebuyers. Various alternative mortgage designs sought to deal with the "tilt problem" created by inflation: specifically when nominal house prices rise over time, a standard mortgage then causes the purchaser to have a higher real value of mortgage payments in the earlier years, resulting in potential cash flow problems for homeowners that will increase the risk of default. Several mortgage designs (with many variants) were developed that led to mortgages with systematically higher real payments over time, including graduated payment mortgages, growing equity mortgages, tiered payment mortgages, shared appreciation mortgages, price-level adjustment mortgages, and dual rate mortgages (Fabozzi and Modigliani 1992).

By the 1980s, the financial sector had certainly recognized that a number of financial risks had increased and needed to be hedged, both in the property sector and from other areas. The property markets in the United States and the United Kingdom became more integrated with their financial markets. In the United

[2]The savings and loan crisis of the 1980s and 1990s was not directly related to changes in property prices. Instead, the key problem was that savings and loan institutions held large portfolios of fixed-rate mortgages, and under existing laws in the early 1980s, they faced a regulatory limit on the interest rate they could pay on deposits. When US inflation and nominal interest rates rose dramatically in the 1970s, savings and loans faced a double-whammy: their deposits flowed away to money market funds, which could pay higher nominal interest rates, and the value of their fixed-rate mortgage assets declined sharply (for discussion, see the three-paper "Symposium on Federal Deposit Insurance" in the Fall 1989 issue of this journal). However, the savings and loan debacle does illustrate a case where issues in the mortgage finance industry led to development of financial derivatives. The development of the interest rate swap market in the 1980s—which made it much easier to exchange fixed-rate and variable-rate securities—is primarily attributable to the need to manage interest risk emerging from the fixed-rate mortgage loan portfolios that were common in the 1970s and earlier.

Kingdom, for example, the Housing Act of 1980 introduced a right-to-buy policy that transformed the UK residential market from a majority of renters to a majority of houseowners. Then 1983 brought the "Big Bang" deregulation of the financial sector (Coakley 1994) at a time when UK bank systems became the custodians of large portions of property risk through large mortgage origination programs.

Many futures contracts were introduced on exchanges throughout the world in the 1980s, including those for the purpose of managing the risks associated with various types of assets from commodities to stock indices and Treasury bonds. But somehow, there were no attempts to introduce futures related to the price of property, whether in the form of residential houses or commercial property. One futures contract of this time was tangentially related to real estate: the Government National Mortgage Association (Ginnie Mae) Collateralized Depositary Receipt (CDR) futures contract. However, this contract focused on interest rates, not property prices. The main users of this contract were mortgage bankers who were holding large undiversified mortgage portfolios which they intended to resell in the secondary market. This financial contract was a modest success for a few years, but then the bulk of investors interested in taking positions on interest rates shifted to a futures contract based directly on US Treasury bonds instead. The story of the six-year rise and fall of the GNMA CDR interest rate futures contract is told in Johnston and McConnell (1989).

However, another property derivative from this time would last longer. Due to the option granted to borrowers to prepay their loan in whole or in part at any time and without penalty, there was considerable uncertainty about what the actual maturity of a mortgage pass-through security might end up being. This "prepayment risk" could result in a security with either a very short maturity or an extremely long maturity. Prepayment risk made mortgage pass-through securities unappealing to traditional investors. In the early 1980s, a new type of mortgage-backed security was created called a collateralized mortgage obligation (CMO) to deal with prepayment risk and the uncertainty of maturities. A CMO was made up of different bond classes (popularly referred to as tranches) backed by a pool of mortgage loans, and it had a set of rules for the distribution of the interest and principal payments to the different bond classes. The rules were such that some of the bond classes carried more prepayment risk than others. When issued by Ginnie Mae, Fannie Mae, and Freddie Mac, there was no concern with credit risk. Other entities also issued CMOs, referred to as private-label CMOs, where the different bond classes had a different credit rating and there were rules not only for the distribution of the interest and principal payments but the allocation of losses to the different bond classes. Overall, these securities resulted in a redistribution of credit risk as well as prepayment risk. The wide range of risk profiles made these securities more appealing to a wide range of institutional investor seeking targeted risk profiles. Unfortunately, it was private-label CMOs backed by a pool of mortgage loans consisting of borrowers with impaired credit ratings (that is, subprime borrowers), which were a main part of the story behind the subprime mortgage crisis of 2007–2008 (Fabozzi 2015, Chapter 11).

The volatility of real house prices had started increasing substantially after the 1970s: one principal reason is the strong increase in residential land prices following World War II (Knoll, Schularick, and Steger 2017). From the late 1980s up to the 2007 subprime crisis and the Great Recession, the rate of growth of real house prices was significantly faster than the rate of income growth.[3] However, this rise has been unequally distributed across locations. Metcalf (2018) reports the changes in the real median house prices for the core-based areas in the United States between 1996 and 2016. The percentage increase varied from 16 percent in Atlanta-Sandy Springs-Roswell to 75 percent in New York-Newark-Jersey City and a maximum of 168 percent in San Francisco-Oakland-Hayward. Thus, many households found themselves in a situation where housing equity represented a large proportion of their personal wealth and where housing equity also seemed like an asset with a degree of risk it would be unwise to disregard.

From the standpoint of the financial sector and the economy as a whole, the total wealth tied up in real estate is extremely high in all developed economies. Since 1870, for the majority of developed economies, the banking sector has gradually moved from business loans to mortgage loans, particularly after World War II. In western countries, total mortgage loans outstanding have risen (on average) from about 20 percent of annual GDP at the beginning of the 20th century, increasing to 70 percent of GDP by 2010 (Jordà, Schularick, and Taylor 2015). The value of US real estate owned by households and nonprofits (that is, not counting property owned by firms) is approximately $30 trillion (Federal Reserve Board of Governors 2019), approaching the value of the US equity market. The estimated value of all developed real estate worldwide, including residential, commercial, and agricultural land is $217 trillion (Savills 2016). However, in 2014, considering all futures and options contracts traded at 78 exchanges around the world, the contracts targeting the property cash market are counted in the "other" category, representing less than 1.4 percent of all derivatives traded (according to the Futures Industry Association website at http://www.fia.org). Hence, there is a clear mismatch between property's market value and the existing property derivatives' notional amount. In addition, there is continued uncertainty as to how well the economies of high-income countries will survive another large, risky event associated with real estate markets. The need to build and strengthen markets in property derivatives is clear.

[3]One possible explanation proffered by Glaeser, Kolko, and Saez (2001) is urbanization. However, that explanation may be more true in some countries than others. For example, for the United States in 1900, 30 percent of the population resided in cities, increasing by 2010 to 80 percent of the population. In Germany in 1910, 60 percent of the population resided in cities, increasing to 75 percent by 2010. In contrast, in the United Kingdom, the cities were occupied by 77 percent of the population in the early 20th century and remained at that same level (approximately 79.5 percent) in 2010 (United Nations 2014; US Bureau of the Census 1975; General Register Office 1951).

Evolution of Real Estate Derivatives

Early Failures and Baby Steps

The first attempt to introduce a standardized house price futures contract occurred in August 1990 when Karl Case, Robert Shiller, and Allan Weiss, under the umbrella of the Case Shiller Weiss Research Group, proposed to the Coffee, Sugar, and Cocoa Exchange a futures market on single-family homes (Shiller 2008). A few months later in November 1990, the Chicago Board of Trade (CBOT) was presented with a similar idea (Jud and Winkler 2009), and Case Shiller and Weiss, Inc. investigated jointly with the CBOT the feasibility of launching a house price futures market. However, a survey in 1993 clearly indicated that the house price market was very one-sided: that is, there were plenty of investors willing to purchase futures contracts to protect themselves against a decline in housing prices, but not so many investors who wanted to sell such contracts, leading the CBOT to decide against launching a house price futures contract at that time.

At almost the same time, the London Futures and Options Exchange (FOX) introduced several property-related futures contracts in May 1991, including a housing futures contract and futures contracts based on prices and rents for a commercial real estate. However, these contracts did not last long. The underlying data series for the house price futures contract was the Nationwide House Price index (NAHP), where the index was constructed using data from home sales on which the Nationwide Anglia Building Society (since 1992, Nationwide) originated mortgage loans (Baum 1991). However, the index became contaminated by unlawful efforts to boost volume by employing "wash trades"—that is, trades in which a single investor is buying and selling equivalent amounts of the contract at the same time, which can be a way to push misleading information into the market. This manipulative practice led to the termination of this contract in October 1991 (Shiller 2008). The commercial property derivatives were terminated at the same time.

In late 1994, the London Futures and Options Exchange attempted to introduce other real estate derivatives based on an index from IPD (Investment Property Databank) but without great success. Barclays de Zoete Wedd introduced Property Index Certificates (PICs) that were later renamed Property Linked Notes because they were effectively euro-bonds that would replicate IPD returns when traded at par (Lizieri et al. 2012).

In 2001, the United Kingdom witnessed the introduction of a betting market based on house prices by the City Index Group and a year later by the IG Index Ltd. Because these markets were perceived as mainly betting opportunities rather than as hedging instruments, trading has been sparse over the years. In May 2005, the Cantor Index, created by a division of Cantor Fitzgerald Group, started offering betting on house prices based on the Average Greater London and Average UK House Price markets.

In the United States, the so-called "hedgelets," promoted by HedgeStreet in October 2004, were futures-type contracts offered online to small investors who had strong convictions on the direction of specific economic indicators (De Aenlloe

2004). These contracts could be used by individuals to make bets in $10 increments on the future direction of house prices. The contracts that individuals could use to bet on the future direction of house prices had a binary or digital characteristic: specifically, the contract was based on whether the housing price index from the Office of Federal Housing Enterprise Oversight (OFHEO) in one of six cities would fall into a given range on a specific date over the next three months. Such a contract implied that if the index failed to fall within the designated range, half of the participants lost their entire investment.

The Arrival of House Price Futures and Options

The first lasting house price futures contract finally arrived on May 22, 2006, when the Chicago Mercantile Exchange (CME) started trading house futures contracts and options based on the family of S&P/Case-Shiller® Home Price Indices, which covered both a national composite index and 10 major cities.[4] This initial contract was a joint collaboration of the CME and MacroMarkets LLC. In February 2008, Standard & Poor's acquired the S&P/Case-Shiller Home Price Indices from MacroMarkets LLC.

For the US commercial real estate market, Standard & Poor's and Global Real Analytics/Charles Schwab Investment Management constructed the S&P/GRA Commercial Real Estate Indices (Labuszewski and Souza 2007), which were then used by the Chicago Mercantile Exchange in November 2007 as the underlying basis for a futures contract. The intention was to trade commercial property futures on the office, warehouse, apartment, and retail property sectors, and more widely for the nation (as well as for the Northeast, Midwest, mid-Atlantic South, Pacific West, and Desert Mountain West regions) with electronic trading out 20 quarters. Trading volume in the S&P/GRA commercial property index futures has been very low. This is probably attributable to the diversity of commercial property indices in the United States, given that there are many indices constructed in different ways, all competing for the interest of market participants.

In the United Kingdom, only one commercial real estate index was recognized by market participants: the family of commercial property indices published by the IPD. The trading of total return swaps on various IPD country indices started the over-the-counter market in January 2005. According to Jud and Winkler (2009), total return swaps were also traded where the underlying was a commercial property index for the countries of Australia, Canada, France, Germany, Italy, Japan, Spain, and Switzerland, with about £17.3 billion (notional value) of swaps referencing the IPD UK index.

One of the most successful property derivatives so far has been futures contracts on the IPD family of commercial real estate indices traded on EUREX in London

[4]The 10 cities with their initial weighting in the composite index are Boston (7.4 percent), Chicago (8.8 percent), Denver (3.6 percent), Las Vegas (1.4 percent), Los Angeles (21.1 percent), Miami (4.9 percent), New York (27.2 percent), San Diego (5.5 percent), San Francisco (11.7 percent) and Washington, DC (7.8 percent).

(Fabozzi and Tunaru 2017). One possible explanation for the relative success of this contract is the fact that the IPD family of indices was and still is widely regarded as the main representative index family for commercial real estate in the United Kingdom. The IPD index construction methodology was extended to other countries such as Germany and France. (By contrast, the multitude of US commercial real estate indices may be detrimental to the innovation of new derivatives financial instruments.) The contracts as initially launched in February 2009 were annual contracts based on the total returns of the IPD UK Annual All Property index for individual calendar years. There are futures related to various property portfolios covered by IPD such as the composite level (UK All Property), sector level (UK office, UK retail, and UK industrial), and sub-sector level (UK retail warehouse, UK shopping centres, London city offices, London west end and mid-town offices, and south-east industrial). One important change that has occurred is the 2015 takeover by Morgan Stanley Capital International (MSCI) of the IPD. MSCI then changed the underlying IPD UK Annual Return All Property index into a quarterly calculation. This change was made to streamline the marking-to-market process to be more in line with the dynamics of the property index.

The Subprime Crisis and a Mortgage Derivative on House Price Risk

Collateralized mortgage obligations were created in the 1980s, as noted in the earlier discussion. These securities had often included a mixture of mortgage of different risk characteristics. However, as a rise in the issuance of subprime mortgages was accompanied by the run-up in housing prices in the early 2000s, there was a sharp increase in the number of investors who were willing to take one side or the other of the market for mortgage risk. This turn triggered the introduction of ABX.HE indices (the initials stand for "asset-backed securities, housing equity").

These indices, which are determined from 20 subprime mortgage-backed securities, employ credit default swaps (CDS). A CDS is an agreement whereby the buyer of protection makes a payment (called a premium) at a regular frequency to the seller of protection. In exchange for the premium, the seller provides some form of price protection for some reference entity over a specified time period should a credit event (such as bankruptcy) occur. CDS contracts initially provided protection for corporate bonds and sovereign bonds, allowing the pricing of credit risk for these issuers. In January 2006, Markit Group, Ltd. introduced the ABX.HE indices. Each index tracks the CDS contracts on subprime mortgages with a specified credit rating at the time the mortgages were originated and issued at a specified time (referred to as the "vintage"). For example, ABX.HE BBB tracked the CDS contracts for subprime mortgages that received a credit rating of BBB. With the introduction of the ABX.HE indices investors were able to obtain transparency about the price of subprime mortgage-backed securities by credit rating. Fender and Scheicher (2008) describe in more detail how changes in the price of the ABX.HE can be interpreted as a barometer for stress in the subprime mortgage market.

The main risk posed by a credit default swap is counterparty credit risk, which is the risk that the seller of protection will not cover the losses in case of a credit event. This type of risk appears not to have been on the radar of regulators prior to the global financial crisis, but it was managed among big market players through collateral posting. Regulators came to recognize this problem when a subsidiary of American International Group (AIG Financial Products) lost almost $100 billion in 2008 alone (for a more in-depth discussion of the AIG story in this journal, see McDonald and Paulson 2015).

On the positive side, the evolution of prices in the ABX.HE market confirms the important role that derivatives contracts can play in providing forward-looking information. The contracts were issued twice a year, in January and June, based on the securities issued in the preceding months. Starting in 2007 and 2008, for example, the prices of contracts issued in June 2006 started falling sharply compared to those issued in January 2006—thus showing that the risk of default on subprime mortgages was rising sharply. Conversely, the prices on the ABX.HE contracts in 2009 signalled the end of the subprime crisis.

Successes and Limitations of the Early Property Derivative Efforts

The combination of the rise and fall in housing prices, the crisis in subprime lending, and the Great Recession, taken together, hobbled the promise of the early direct hedging vehicles for real estate risks. The size of the futures and options markets for the S&P/Case-Shiller Home Price Index on the Chicago Mercantile Exchange peaked around the time of the subprime crisis and survives only in a diminished form. In London and other cities, the IPD swaps market grew dramatically until around 2008 but have languished since then. However, in 2009 Eurex launched futures contracts on several IPD indices for various sectors of the commercial real estate market which are still traded today. Overall, the UK property derivatives market has experienced more success than its US counterpart. Torous (2017) offers two possible explanations: 1) the UK market has one dominant commercial real estate index while the US market has several; and 2) there has been effective lobbying by UK property funds to adopt to new more favorable tax legislation.

Studies of the early efforts at creating property derivatives have clearly demonstrated their potential benefits, for example. Lee, Stevenson, and Lee (2014) and Wong, Chau, and Yiu (2007) provide empirical evidence on the stabilizing role of property futures on the volatility of spot property markets. Zhu, Pace, and Morales (2014) empirically investigate how well market information from the Case-Shiller house price futures performed as a forward-looking forecast. Using loan-level mortgage data covering over 90 percent of the residential mortgage loans included in the mortgage pool of US non-agency securitized deals, they found that forecasts extracted from the Case-Shiller house price futures outperformed other proxies preferred in the literature and employed in practice, both in sample and out of sample. Moreover, the Case-Shiller futures forecasts were the only series implying a downward housing price effect that would impact negatively on mortgage default behavior.

Property derivatives provided institutional investors, such as pension funds and insurance funds, a tool to manage their commercial real estate portfolio more efficiently. Bertus, Hollans, and Swidler (2008), for example, demonstrate that investors exposed to house price risk in Las Vegas could have hypothetically have used the CME house price index futures to reduce risk by more than 88 percent from 1994 to 2006 (one year prior to the subprime crisis). Information extracted from the price of property derivatives can play an important role in providing expectations of housing prices that can be used in modelling mortgage defaults. Dolan and Hume (2010) show that the CME futures market effectively predicted the home price crash in the United States before the news media did.[5] Jud and Winkler (2008, 2009) look at risk and return for an investor who participated in the house price futures market. Using daily data on CME-traded house price futures for the period May 2006–May 2008, they reported that the returns on futures were positive, even if the returns of investing in the spot market were negative.

Empirical evidence covering a few European countries including the United Kingdom also highlights the substantial benefits associated with house price derivatives when utilized to manage risk (Englund, Hwang, and Quigley 2002; Iacoviello and Ortalo-Magné 2003; Quigley 2006). These benefits include increasing the financial system's stability, and the ability of millions of homeowners to manage property risk more cost-effectively (Fabozzi, Shiller, and Tunaru 2009). Bond and Mitchell (2011) also find that property derivatives prices outperformed the consensus forecasts of future returns in the UK market.

Obstacles in the Development of Residential Property Derivatives

Several surveys of key players in real estate markets have inquired about the reasons for their reluctance to trade property derivatives.[6] Here, we focus on what we see as some of the most prominent impediments to growing a market for financial instruments based on house prices: 1) how real estate indexes may be mismatched with the needs of property derivatives; 2) a fear of negligible liquidity; 3) the lack of models to price these derivatives; and 4) concerns about an uncertain legislative framework vis-à-vis this new derivatives asset class. Along with the four concerns discussed here about hindrances to a more robust development of property derivatives, other concerns mentioned in the surveys include a lack of education by house

[5]John H. Dolan, market maker for eight years for the CME Case Shiller home price futures and options markets, has a web site, *HomePriceFutures.com,* that provides regular information about those markets,and moderates an online discussion on LinkedIn (the "CME Case Shiller Home Price Futures" group).

[6]For example, Lim and Zhang (2006) use a web-based survey of 37 US-based real estate investment managers, fund managers, and commercial lenders and brokers to identify the principal reasons for the stalling development of property derivatives. Venter (2007) interviews ten UK individuals that included tax lawyers, an index provider, investment bankers, brokers, investment advisors, and a property company. Puntener (2011) interviews six academic experts, 17 financial and property experts, and two advisors, in the United States and United Kingdom. Hanisch (2019) carries out 41 individual interviews and two group interviews between June 2016 and March 2017.

owners on appropriate use of derivatives and the number of asset managers who see little need to hedge real estate risk because of the low volatility traditionally associated with sectors of the real estate market. There was no evidence of regulatory or cost barriers that may have deterred potential entrants into the property derivatives sector prior to the subprime crisis. However, this has changed in the aftermath of the subprime crisis and some stringent regulatory risk measures have been imposed on derivatives in general.

Index Construction Mismatch

The construction of a house price index involves a number of choices, which in turn affects the financial derivatives that might be built using such an index. For example, house price indices can be national or regional, rural or urban, cover new or existing homes, or both. Prices for an index might be taken from market house sales, which runs the risk of not representing properties with the same characteristics over time. Alternatively, a house price index might use listed prices (whether or not a sale results) or appraisals by designated organizations, but these possibilities include more subjectivity on what a certain property is "worth."

The most widely preferred method is to construct real estate indexes by using sales prices, but in a way that adjusts for the quality of the houses being sold. One approach is to use hedonic regressions, thus adjusting for key qualities of the house. However, real estate has a wide array of unobserved heterogeneity, including aspects of location and quality of maintenance and upkeep, so there is reason to doubt that the assumptions made when using a hedonic approach are satisfied (Clapham et al. 2006). Thus, real estate indexes have mostly converged on the repeat-sales approach, which focuses on houses that have been sold at two different points in time. Case and Shiller (1987) propose the weighted repeat-sales method: in their version, repeat sales that happen with a longer time interval between sales are given less weight than repeat sales with smaller time intervals, because the quality of a house changes more over longer time periods. The S&P/Case-Shiller® Home Price Indices use this approach, and the Office of Federal Housing Enterprise Oversight (OFHEO) publishes a repeated sales index using its own version of the Case-Shiller approach.

The problem arises because a standard futures contract, like the CME house price futures contract, is based on the initial value of the real estate index (CME 2007a, b). Over the life of the futures contract, new information is gathered on contemporaneous paired sales. This new information most likely will create changes in the estimates of the house price index value in all previous periods. A large discrepancy can arise between futures settled on the initial value of the index and those settled on the revised value of the spot index. This potential discrepancy is sensitive to details of how the index is weighted and calculated (Shiller 1993a, b; Deng and Quigley (2008).

Problems in matching the timing of the real estate index to the property derivative can arise in a number of ways. In the United Kingdom, property futures traded on the IPD exchange use a December year-end maturity for five years ahead, but

a post-March publication date for the real estate index: hence, there is a three-month period mismatch between the calculation period (December to December) and the information period (March to March). Another design problem for the IPD futures contracts was that the futures were traded on the IPD UK Annual total return index, while the marking-to-market was done on the IPD UK Monthly total return index. The latter index covered only a subsample of about 80 percent of the annual IPD index. The acquisition of IPD by the MSCI in 2012 had implications for the commercial real estate index family: the major change involved switching to a quarterly valuation. As another example, the existing MSCI UK Quarterly property index captures total returns of directly held standing property investment, based on tracking the performance of 8,913 property investments, cumulating to £160 billion by June 2019.

The ties between calculation of real estate indexes and the property derivatives based on those indexes are real ones. For example, using an extensive repeat-sales database for the Paris housing market, Baroni, Barthélémy, and Mokrane (2008) show that the revision problem may cause concern about the stability of some key parameters. However, the magnitude of the impact of revision on the property price indexes is not so substantial so as to make market participants pull out of property derivatives that would protect them against severe market downturns.

Negligible Liquidity: Missing One Side of the Market?

Most real estate owners recognize that they have made a long-term investment in an asset they will some day wish to sell at the spot price, and so they are at least potentially interested in property derivatives to hedge against the risk of falling prices. In addition, Jordà, Schularick, and Taylor (2015) show that in developed economies, by 2010, mortgage credit on the balance sheets of banks represented about 60 percent of assets on average. Moreover, the stress testing that has been introduced for systemically important banks and financial institutions requires those entities to pass an overall portfolio survival test against a decline of 30 percent in real estate markets. One way to satisfy the regulators would be to purchase an option on the major real estate indices that would only pay off after a substantial decline in property values. If such an option was traded regularly, liquidity in the property derivative market would receive a welcome boost. In general, futures contracts have provided a reliable vehicle to offset risk in capital markets. For property markets, futures contracts also allow investors to take positions that are equivalent to short-selling the property market, which is not possible to do in the spot markets for property.

But at first glance, it is unclear who should be the counterparty in those property derivatives trades: that is, who is willing to provide insurance against a fall in property prices or hedge against a rise in prices? Of course, a property derivatives market cannot flourish without participants on both sides. Any investor who has exposure to a drop in property prices should be interested in offsetting possible losses on their exposure with the financial gains from a position in property derivatives. This is the case for typical homeowners, real estate financial houses, institutional investors in

mortgage-backed securities, pension funds invested in property portfolios, insurance funds using property investments for their asset-liability management, and building societies who carry mortgages on their balance sheet.

One can conjure up hypothetical examples for the other side of the market easily enough: as mentioned earlier, one can imagine young people, who expect to buy their first home when they are ready in the future, might use property derivatives to start investing in property synthetically to avoid being priced out of the market. Homeowners in one city who feel they will eventually move to some other city might combine a short position, that is, selling the futures contract in their current city, with a long position, that is, buying the futures contract in a national home index price. Providers of "target date" retirement funds might provide such a service, perhaps adjusting exposure to real estate risk in the local market and in likely retirement destinations as the beneficiary approaches retirement age. However, it seems implausible that these kinds of market participants will be substantial enough to make up the other side of the property derivatives market.

More likely, it would be mutual funds, insurance companies, pension funds and other managers of large pools of funds who desire to be fully diversified who take the other side of the real estate risk on derivatives. The role of speculators and sophisticated traders, such as hedge funds and private equity firms, becomes even more important to ensure liquidity for property derivatives. Once a market in property derivatives is clearly established, it's also easy to imagine that general investors might eventually, after a market is established, be enticed by how the combination of risks and returns fits within their broader portfolio.

Modelling Considerations

Given the non-standard characteristics of real estate indexes and property in general by comparison to commodities, equities, bonds, and currency exchange rates, it is perhaps not surprising that pricing even straightforward derivatives such as futures, put and call options, and total return swaps is not straightforward. Many pitfalls and caveats must be considered. For example, many of the models used for pricing derivatives depend on a no-arbitrage constraint: thus, the futures market for the S&P 500 as a whole is governed by what it would cost to buy the portfolio of underlying stocks. But buying a portfolio of houses that replicates a well-designed real estate index would be a costly and illiquid investment, so this no-arbitrage condition and its implications no longer hold. A standard no-arbitrage condition suggests that the relationship between the spot price of the derivative and the expected future price will be driven largely by the risk-free rate of return, but as Drouhin, Simon, and Essafi (2016) show in a study of IPD total return swaps contracts, this relationship does not hold in the context of property prices. Furthermore, without a no-arbitrage condition, the standard Black-Scholes option-pricing formula cannot be derived using the classical replication approach.

Indeed, one argument for the full development of a property derivatives market with futures contracts is that it sidesteps what can otherwise be some complex and disputable econometric work on valuing derivatives. If property

futures do exist in the market, then it becomes possible to set prices for options on property prices using the Black (1976) formula. However, market-makers still need to decide those prices for futures somehow. While producing different valuations of property derivatives may attract more players into this market, market-makers need to make defensible decisions that are capable of resisting attempts by various other players to drive the market price too high or too low for non-economic reasons. Models that can be relied upon for property derivatives markets are also models that take into account the econometric features of the spot property markets. They must be easy to set up, not highly computationally intensive, and characterized by parameters that have a direct interpretation in financial terms.

There are currently two schools of thought about the appropriate models to price property derivatives. One school considers how to replicate the contingent claim of the property derivative given existing prices in the market, along with selecting a set of stochastic processes to represent dynamics of the underlying real estate index combined with risk-neutral pricing. Examples along these lines include Titman and Torous (1989), Buttimer, Kau, and Slawson (1997), Björk and Clapham (2002), Otaka and Kawaguchi (2002), Syz and Vanini (2011), and Fabozzi, Shiller, and Tunaru (2012).

A main challenge in this approach is that the choice of stochastic processes to represent dynamics of the underlying real estate index can lead to difficulties. One unreliable approach followed in some strands of the literature assumed that the underlying property indices or property prices followed a geometric Brownian motion. This assumption is inconsistent with the overwhelming empirical evidence starting with Case and Shiller (1987, 1989) that indicates that house prices 1) exhibit serial correlation and 2) are positively correlated over short horizons and negatively correlated over long horizons. More recent evidence across several markets is presented in Tunaru (2017). Mean-reverting processes are capable of generating pathways that match these empirical characteristics and they could be more appropriate as a starting point for pricing property derivatives as discussed in more detail in Fabozzi, Shiller, and Tunaru (2012).

The other school of thought in this literature is defined by various equilibrium models. For example, Geltner and Fisher (2007) and Lizieri et al. (2012) propose equilibrium-based models for calculating forward prices and the total return swap spread. Cao and Wei (2010) sidestep the non-tradability of underlying housing indices for the CME-traded housing futures and options by assuming a mean-reverting aggregate dividend process and a constant relative risk aversion utility function to derive analytical forwards and options prices, equilibrium, and no-arbitrage.

However, this approach also raises questions. Equilibrium models may be useful for marking-to-model property derivatives positions, particularly when there is no information on the derivatives either due to market closure or crises events. However, the connections between a known futures price given by markets and corresponding prices of put and call options are based on model-free, no-arbitrage

relationships, and Tunaru (2017) provides several examples showing that equilibrium prices may not satisfy this requirement. Perhaps the biggest concern is that if one of the arguments for property derivatives is as a mechanism to foresee and to manage financial crises, it is difficult to reconcile the idea of a market being in equilibrium and in a financial crisis at the same time.

A final concern about modelling futures prices in property derivatives is that the market for single-family homes has been one of exceptionally high transactions cost and impossibility of short sales, which permit the high level of momentum and of apparent bubbles. The transition to a real estate market with functioning property derivatives that allow for extensive hedging may well alter the time-series properties of the underlying cash price. For that reason, the past time-series properties of home price indices may not be a good guide to the future. However, one can view this as a transition problem, which would become smaller over time as market experience increases with property derivatives.

Regulatory Issues

Before the subprime crisis, a number of large investment banks were involved in property total return swaps using over-the-counter trades—for example, Deutsche Bank, Merrill Lynch, Morgan Stanley, and the Royal Bank of Scotland. They were satisfied to enter trades with various clients and take the risk on their books for long periods of time until they were able to offload those risks.

In the aftermath of the subprime crisis, trading in the property derivatives market moved from over-the-counter to exchange-based. In addition, the Basel III Accord established a new set of rules requiring banks to allocate additional capital for each leg of a derivatives trade. As a result, trading property derivatives became very capital intensive. These increased regulatory capital requirements associated with property derivatives motivated banks and investment banks to exit this asset class. Given ongoing concerns about financial risk, bailouts, and systemic risk, there is ongoing concern about future rules that might further discourage large financial institutions from participating in property derivatives markets.

Lessons and Proposals for the Future

Since the 1970s, property price risk has affected investors and economies with increased frequency. Markets in property derivatives are the key to providing both investors and lenders with the tools to mitigate property-related risks. However, the market for real estate has various characteristics that differentiate it from other asset classes where derivatives were successfully introduced. Given the specific economic and econometric characteristics of the underlying asset, along with the house price and commercial property indexes based on that asset, property derivatives require a more complex process to be generally accepted by financial market participants. In particular, more needs to be done on the modelling side to facilitate pricing and hedging in this incomplete market. The ultimate goal is for property derivatives

to be traded as a standard commodity, similar to the way that futures, options, and swaps are traded for stock and bond indexes.

Financial derivatives have sometimes attracted a bad reputation, often after prominent financial institutions (like the AIG example with credit default swaps mentioned earlier) suffered large losses. Derivatives can allow for high leveraging and when events turn bad, may magnify losses. In modern times, the majority of financial crises involved in one way or another the use, or rather, the misuse of derivatives. Of course, with or without financial derivatives, investors have many ways to underestimate risk and end up with substantial losses. In contrast, during the many times when financial derivatives have allowed parties to hedge risk, increase speed, reduce transactions costs, and balance investment portfolios, it has attracted almost no attention. By now, derivatives are commonly used worldwide, and their usefulness in spreading various types of risk in a sustainable manner is gradually passing the test of time.

Governments, banks, and other financial institutions have sound reasons to work together to give impetus to the development of property derivatives. As the largest asset class without corresponding liquid derivatives, property derivatives would offer some of the largest benefits from making the leap to a commoditization status. This leap would help directly and indirectly provide forward-looking price signals for a variety of uses, including their application to stabilizing financial systems, and in this way, reduce the risk of market crashes and the resulting economic instability.

The historical development of derivatives markets to deal with the risks of other asset classes such as equity, foreign exchange, bonds, commodities and credit default swaps, suggests that those derivatives markets were greatly helped by a model that was generally adopted by the important market players—the Black-Scholes (1973) model for valuing equity options being the most notable example. As market volume increased, so did the demand for innovation in those markets that led to the introduction of more sophisticated models. But currently, property derivatives lack a widely accepted model.

References

Baroni, Michel, Fabrice Barthélémy, and Mahdi Mokrane. 2008. "Is it Possible to Construct Derivatives for the Paris Residential Market?" *Journal of Real Estate Finance and Economics* 37: 233–64.

Baum, Andrew. 1991. "Property Futures." *Journal of Property Valuation and Investment* 9 (3): 235–40.

Bertus, Mark, Harris Hollans, and Steve Swidler. 2008. "Hedging House Price Risk with CME Futures Contracts: The Case of Las Vegas Residential Real Estate." *Journal of Real Estate Finance and Economics* 37: 265–79.

Björk, Tomas, and Eric Clapham. 2002. "On the Pricing of Real Estate Index Linked Swaps." *Journal of Housing Economics* 11 (4): 418–32.

Black, Fischer. 1976. "The Pricing of Commodity Contracts." *Journal of Financial Economics* 3 (1–2): 167–79.

Black, Fischer and Myron Scholes. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81: 637–654.

Bond, Shaun A., and Paul Mitchell. 2011. "The Information Content of Real Estate Derivative Prices." *Journal of Portfolio Management* 37 (5): 170–81.

Buttimer, Richard J., James B. Kau, and V. Carlos Slawson Jr. 1997. "A Model for Pricing Securities Dependent upon a Real Estate Index." *Journal of Housing Economics* 6 (1): 16–30.

Cao, Melanie, and Jason Wei. 2010. "Valuation of Housing Index Derivatives." *Journal of Futures Markets* 30 (7): 660–88.

Caplin, Andrew, William N. Goetzmann, Eric Hangen, Barry J. Nalebuff, Elisabeth Prentice, John Rodkin, Matthew I. Spiegel, and Tom Skinner. 2003. "Home Equity Insurance: A Pilot Project." Yale International Center for Finance Working Paper 03–12.

Case, Karl E., and Robert J. Shiller. 1987. "Prices of Single Family Homes since 1970: New Indexes for Four Cities." *New England Economic Review* September: 45–56.

Case, Karl E., and Robert J. Shiller. 1989. "The Efficiency of the Market for Single-Family Homes." *American Economic Review* 79 (1): 125–137.

Case, Karl E., and Robert J. Shiller. 1996. "Mortgage Default Risk and Real Estate Prices: The Use of Index-Based Futures and Options in Real Estate." *Journal of Housing Research* 7 (2): 243–58.

Case, Karl E., Robert J. Shiller., and Allan N. Weiss. 1991. "Index-Based Futures and Options Markets in Real Estate." Cowles Foundation Discussion Paper 1006.

Case, Karl E., Robert J. Shiller., and Allan N. Weiss. 1993. "Index-Based Futures and Options Markets in Real Estate." *Journal of Portfolio Management* 19 (2): 83–92.

Clapham, Eric, Peter Englund, John M. Quigley, and Christian L. Redfearn. 2006. "Revisiting the Past and Settling the Score: Index Revision for House Price Derivatives." *Real Estate Economics* 34 (2): 275–302.

CME. 2007a. "CME Housing Future and Options: Opening up New Opportunities." *Chicago Mercantile Exchange.*

CME. 2007b. "CME Housing Futures and Options: Frequently Asked Questions." *Chicago Mercantile Exchange.*

Coakley, Jerry. 1994. "The Integration of Property and Financial Markets." *Environment and Planning A* 26 (5): 697–713.

De Aenlle, Conrad. 2004. "A New Way to Hedge Your Home's Paper Profit." *The New York Times.* December 12.

De Long, J. Bradford, Andrei Shleifer, Larry H. Summers, and Robert J. Waldmann. 1990. "Noise Trader Risk in Financial Markets." *Journal of Political Economy* 98 (4): 703–38.

Deng, Yongheng, and John M. Quigley. 2008. "Index Revision, House Price Risk, and the Market for House Price Derivatives." *Journal of Real Estate Finance and Economics.* 37: 191–209.

Dolan, John H. and Susan R. Hume. 2010. "Observations on the CME Home Price Futures Market: Were these Futures Able to Predict the Home Price Crash?" *Proceedings of the Northeast Business & Economics Association* 269–73.

Drouhin, Pierre-Arnaud, Arnaud Simon, and Yasmine Essafi. 2016. "Forward Curve Risk Factors Analysis in the UK Real Estate Market." *Journal of Real Estate Finance and Economics* 53: 494–526.

Englund, Peter. 2010. "Trading on Home Price Risk: Index Derivatives and Home Equity Insurance." In *The Economics of Housing: The Housing Wealth of Nations,* edited by Susan J. Smith and Beverley A. Searle, 499–511. Oxford: Wiley-Blackwell.

Englund, Peter, Min Hwang, and John M. Quigley. 2002. "Hedging Housing Risk." *Journal of Real Estate Finance and Economics* 24: 167–200.

Fabozzi, Frank J. 2015. *Capital Markets: Institutions, Instruments, and Risk Management.* 5th ed. Cambridge, MA: MIT Press.

Fabozzi, Frank. J., and Franco Modigliani. 1992. *Mortgage and Mortgage-Backed Securities Markets.* Boston: Harvard Business School Press.

Fabozzi, Frank J., and Radu S. Tunaru. 2017. "Commercial Real Estate Derivatives: The End or the Beginning?" *Journal of Portfolio Management* 43 (6): 179–86.

Fabozzi, Frank J., Robert J. Shiller, and Radu S. Tunaru. 2012. "A Pricing Framework for Real Estate Derivatives." *European Financial Management* 18 (5): 762–89.

Fabozzi, Frank J., Robert J. Shiller, and Radu S. Tunaru. 2019. "Evolution of Real Estate Derivatives and Their Pricing." *Journal of Derivatives.* 26 (3): 7-21.

Federal Reserve Board of Governors. 2019. *Financial Accounts of the United States. Flow of Funds, Balance Sheets and Integrated Macroeconomic Accounts.* Washington, DC: Board of Governors of the Federal Reserve System.

Fender, Ingo, and Martin Scheicher. 2008. "The ABX: How Do the Markets Price Subprime Mortgages Risk." *BIS Quarterly Review* September: 67–81.

Geltner, David, and Jeffrey D. Fisher. 2007. "Pricing and Index Considerations in Commercial Real Estate Derivatives." *The Journal of Portfolio Management* 33 (5): 99–118.

Gemmill, Gordon. 1990. "Futures Trading and Finance in the Housing Market." *Journal of Property Finance.* 1 (2): 196–207.

General Register Office. 1951. *Census 1951, England and Wales: Preliminary Report* London: HMSO.

Gertler, Mark, and Simon Gilchrist. 2018. "What Happened: Financial Factors in the Great Recession." *Journal of Economic Perspectives* 32 (3): 3–30.

Glaeser, Edward L. 2013. "A Nation of Gamblers: Real Estate Speculation and American History." *American Economic Review* 103 (3): 1–42.

Glaeser, Edward. L., Jed Kolko, and Albert Saiz. 2001. "Consumer City." *Journal of Economic Geography* 1 (1): 27–50.

Goetzmann, William M., and Frank Newman. 2010. "Securitization in the 1920s." NBER working paper 15650.

Goodman, Laurie S., and Christopher Mayer. 2018. "Homeownership and the American Dream." *Journal of Economic Perspectives* 32 (1): 31–58.

Hanisch, Alexander T. 2019. "Factors Influencing the Propensity of Real Estate Investors in the UK to Employ Property Derivatives." *Journal of Property Investment and Finance* 37 (2): 194–214.

Himmelberg, Charles, Christopher Mayer, and Todd Sinai. 2005. "Assessing High House Prices: Bubbles, Fundamentals and Misperceptions." *Journal of Economic Perspectives* 19 (4): 67–92.

Iacoviello, Matteo, and François Ortalo-Magné. 2003. "Hedging Housing Risk in London." *Journal of Real Estate Finance and Economics* 27: 191–209.

Johnston, Elizabeth Tashjian, and John J. McConnell. 1989. "Requiem for a Market: An Analysis of the Rise and Fall of a Financial Futures Contract." *Review of Financial Studies* 2 (1): 1–23.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2015. "Betting the House." *Journal of International Economics* 96 (S): S2–18.

Jud, Donald G. and Daniel T. Winkler. 2008. "Housing Futures Markets: Early Evidence of Return and Risk." *Journal of Housing Research* 17 (1): 1–12.

Jud, Donald G. and Daniel T. Winkler. 2009. "The Housing Futures Market." *Journal of Real Estate Literature* 17 (2): 181–203.

Knoll, Katharina, Moritz Schularick, and Thomas Steger. 2017. "No Price Like Home: Global House Prices, 1870–2012." *American Economic Review* 107 (2): 331–53.

Kothari, Vinod. 2012. "Covered Bonds." In *The Handbook of Fixed Income Securities,* edited by Frank J. Fabozzi, 459–72, New York: McGraw Hill.

Labuszewski, John W, and Kleber L.A. Souza. 2007. "Introduction to CME U.S. Commercial Real Estate Futures and Options." Paper presented at the annual meeting of the American Real Estate Society, San Francisco.

Lee, Chyi, Simon Stevenson, and Ming-Long Lee. 2014. "Futures Trading, Spot Price Volatility and Market Efficiency: Evidence from European Real Estate Securities Futures." *Journal of Real Estate Finance and Economics.* 48 (2): 299-322.

Lim, Jong Yoon, and Yi Zhang. 2006. "A Study on Real Estate Derivatives." https://mitcre.mit.edu/wp-content/uploads/2012/11/real-estate-derivatives.pdf.

Lizieri, Colin, Gianluca Marcato, Paul Ogden, and Andrew Baum. 2012. "Pricing Inefficiencies in Private Real Estate Markets Using Total Return Swaps." *Journal of Real Estate Finance and Economics* 45: 774–803.

McCarthy, Jonathan, and Richard W. Peach. 2004. "Are Home Prices the Next 'Bubble'?" *FRBNY Economic Policy Review* 10 (3): 1–17.

McDonald, Robert, and Anna Paulson. 2015. "AIG in Hindsight." *Journal of Economic Perspectives* 29 (2): 81–106.

Metcalf, Gabriel. 2018. "Sand Castles before the Tide? Affordable Housing in Expensive Cities." *Journal of Economic Perspectives* 32 (1): 59–80.

Mian, Atif, and Amir Sufi. 2018. "Finance and Business Cycles: The Credit-Driven Household Demand Channel." *Journal of Economic Perspectives* 32 (3): 31–58.

Nakajima, Makoto, and Irina A. Telyukova. 2017. "Reverse Mortgage Loans: A Quantitative Analysis." *Journal of Finance* 72 (2): 911–50.

Otaka, Masaaki, and Yuichiro Kawaguchi. 2002. "Hedging and Pricing of Real Estate Securities under Market Incompleteness." Meikai University Working Paper.

Puntener, Sonja Yvonne. 2011. "Determinants of Property Derivatives: Market Architecture and Evolution." Doctoral Thesis. University of St. Gallen.

Quigley, John M. 2006. "Real Estate Portfolio Allocation: The European Consumers' Perspective." *Journal of Housing Economics.* 15 (3): 169–88.

Savills World Research. 2016. Around the World in Dollars and Cents. www.savills.com. (accessed July 20, 2019).

Shiller, Robert J. 1993a. "Measuring Asset Value for Cash Settlement in Derivative Markets: Hedonic Repeated Measures Indices and Perpetual Futures." *Journal of Finance.* 48 (3): 911–31.

Shiller, Robert J. 1993b. "The Theory of Index-Based Futures and Options Markets." *Estudios Economicos.* 8 (2): 163–78.

Shiller, Robert J. 1993c. *Macro Markets: Creating Institutions for Managing Society's Largest Economic Risks.* Oxford: Oxford University Press.

Shiller, Robert J. 2005. *Irrational Exuberance.* 2nd ed. Princeton, NJ: Princeton University Press.

Shiller, Robert J. 2006. "Long-Term Perspectives on the Current Boom in Home Prices." *Economists' Voice* 3 (4): 1–11.

Shiller, Robert J. 2008. "Derivatives Markets for Home Prices." NBER Working Paper 13962.

Shiller, Robert J. 2014. "Why Is Housing Finance Still Stuck in Such a Primitive Stage?" *American Economic Review* 104 (5): 73–76.

Shiller, Robert J., and Allan N. Weiss. 1999. "Home Equity Insurance." *Journal of Real Estate Finance and Economics* 19 (1): 21–47.

Syz, Juerg M., and Paolo Vanini. 2011. "Arbitrage Free Price Bounds for Property Derivatives." *Journal of Real Estate Finance and Economics* 43: 281–98.

Thomas, Guy R. 1996. "Indemnities for Long-Term Price Risk in the UK Housing Market." *Journal of Property Finance* 7 (3): 38–52.

Titman, Sheridan, and Walter Torous. 1989. "Valuing Commercial Mortgages: An Empirical Investigation of the Contingent-Claims Approach to Pricing Risky Debt." *Journal of Finance* 44 (2): 345–73.

Torous, Walter N. 2017. "History of Commercial Real Estate Derivatives." Paper presented at Center for Real Estate, Sloan School of Management, MIT, Cambridge, MA, October 13.

Tsatsaronis, Kostas, and Haibin Zhu. 2004. "What Drives Housing Price Dynamics: Cross-country Evidence." *BIS Quarterly* March: 65–78.

Tunaru, Radu S. 2017. *Real-Estate Derivatives: From Econometrics to Financial Engineering.* Oxford: Oxford University Press.

Uluc, Arzu. 2018. "Stabilising House Prices: The Role of Housing Futures Trading." *Journal of Real Estate Finance and Economics* 56: 587–621.

United Nations. 2014. "Online Data: Urban and Rural Population." United Nations. https://www.un.org/en/development/desa/population/publications/dataset/urban/urbanAndRuralPopulationByAgeAndSex.asp (accessed July 20, 2019).

US Census Bureau. 1975. "Historical Statistics of the United States: Colonial Times to 1970." US Department of Commerce, Washington, DC. https://www.census.gov/library/publications/1975/compendia/hist_stats_colonial-1970.html (accessed July 20, 2019).

Venter, Jani. 2007. "Barriers to Growth in the US Real Estate Derivatives Market." Thesis. Massachusetts Institute of Technology.

Voicu, Cristian, and Michael Joseph Seiler. 2013. "Deriving Optimal Portfolios for Hedging Housing Risk." *Journal of Real Estate Finance and Economics* 46: 379–96.

Wheaton, William C., Mark S. Baranski, and Cesarina A. Templeton. 2009. "100 Years of Commercial Real Estate Prices in Manhattan." *Real Estate Economics* 37 (1): 69–83.

Wong, Kei S., Wing K. Chau, and Yim C. Yiu. 2007. "Volatility Transmission in the Real Estate Spot and Forward Markets." *Journal of Real Estate Finance and Economics.* 35(3): 281-93.

Zhu, Shuang, Kelley Pace, and Walter A. Morales. 2014. "Using Housing Futures in Mortgage Research." *Journal of Real Estate Finance and Economics* 48: 1–15.

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 146–167

Welfare Analysis Meets Causal Inference

Amy Finkelstein and Nathaniel Hendren

Economists have made remarkable progress over the last several decades in developing empirical techniques that provide compelling evidence of causal effects—the so-called "credibility revolution" in empirical work (as discussed in this journal by Angrist and Pischke 2010). But while it is interesting and important to know what the effects of a policy are, we are often also interested in a normative question as well: Is the policy a good idea or a bad idea? Or in the more careful language of economics: What is the welfare impact of the policy?

Until recently, there had been relatively little effort to harness the gains of the "credibility revolution" for the goal of welfare analysis. Instead, we in the empirical public finance community have struggled with other approaches. One venerable tradition is to take an estimate of the benefits of an expenditure policy and compare it to the "cost" to the government. This cost is usually defined as the expenditures on the program, multiplied by 1 plus "the marginal cost of public funds," which is designed to take account of the distortionary effects of the taxation needed to finance the policy, which everyone "in the know" knows to be 0.3, or maybe 0.5 if you're feeling pessimistic. Thus, buried in the last section of an empirical paper that painstakingly estimates the impact of a policy, is an ad hoc analysis that compares the benefits to the cost, multiplied by a smudge factor of 1.3 (for example, Finkelstein and McKnight 2008; Olken 2007). The other common method is the "marginal

■ *Amy Finkelstein is Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Nathaniel Hendren is Professor of Economics, Harvard University, Cambridge, Massachusetts. Their email addresses are afink@mit.edu and nhendren@fas.harvard.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.4.146.

excess burden" or "deadweight loss" approach, which requires valiant attempts to separate non-distortionary income effects of policies (which are transfers rather than welfare losses) from their distortionary substitution effects (which lead to deadweight losses). As Goolsbee (1999) has lamented, "The theory largely relates to compensated elasticities, whereas the natural experiments provide information primarily on the uncompensated effects."

Fortunately, glimpses of light have appeared at the end of the empirical-welfare tunnel. In this essay, we describe a transparent framework for mapping empirical estimates of causal effects of a public expenditure (or tax) change to welfare analysis of that policy change. Following Hendren and Sprung-Keyser (2020), we refer to it as the "marginal value of public funds" (MVPF). The MVPF is the ratio of the marginal benefit of the policy to the net marginal cost to the government of the policy; crucially, this net marginal cost is inclusive of the impact of any behavioral responses to the policy on the government budget.

Our goal is not to break new theoretical ground with the MVPF framework—as we will discuss, its mathematical formulation has been around for decades (for example, Mayshar 1990). Instead, we proselytize for the underrecognized empirical usefulness of this approach in the wake of the "credibility revolution." Its key advantage is that it relies on the causal effects of policy changes to conduct empirical welfare analysis. We provide guidance on how to implement and interpret this approach, with the hope that it will facilitate empirical welfare analysis across a variety of fields.

To do so, we start with a benchmark case of a small increase in a cash transfer that only affects its recipients, whose response to the policy is privately optimal. Under these assumptions, we show that estimates of causal effects of the policy are needed only for estimating the policy's costs, not its benefits. We discuss how this logic is adapted as we relax each simplifying assumption in the benchmark case.

Once we estimate the MVPF, how can we use it? An MVPF of, say, 1.5 means that every $1 of net government spending provides $1.50 of benefits to the beneficiaries of the policy—or in other words, the beneficiaries would be willing to pay up to $1.50 for that $1 policy. One can use the MVPF to compare this "bang for the buck" of policies that affect the same group of individuals (the policy with the higher MVPF is preferable) or for comparing policies that affect different groups. In the latter case, the MVPF quantifies the implicit tradeoffs involved: given policies A and B, policy A is preferred to policy B if and only if one prefers giving $MVPF_A$ to policy A beneficiaries over giving $MVPF_B$ to policy B beneficiaries. Of course, economists have no special powers that allow them to declare such tradeoffs are appropriate.[1] But economics can clarify the tradeoffs embodied in the policy decisions society faces.

[1] In other words, the National Bureau of Economic Research prohibition for its working papers against "statements regarding which policies should (or should not) be adopted" (at http://papers.nber.org/wpsubmit/wp_submit.html) encodes a fundamental and important recognition of the limits of economic analysis.

We then endeavor to answer some common and natural questions about the MVPF approach, including how it relates to "traditional" public finance welfare tools like marginal excess burden and marginal cost of public funds. Finally, we offer some examples of how the MVPF approach has or can be applied to some recent empirical applications across a variety of fields, including public finance, labor economics, development economics, trade, and industrial organization.

How to Construct the Marginal Value of Public Funds: An Initial Illustration

The MVPF is defined as the ratio of the marginal benefit of the policy to the marginal cost of the policy. Equivalently (and more usefully for operationalizing it), it is the ratio of the beneficiaries' willingness to pay for the increase in expenditure out of their own income to the net cost to the government of the increase in expenditure per beneficiary:

$$\text{MVPF} = \frac{\text{Beneficiaries' Willingness to Pay}}{\text{Net Cost to Government}}.$$

Let's consider how to calculate the MVPF for a $1 change in cash benefits in a public program. This could be, for example, a means-tested cash welfare program like Temporary Assistance to Needy Families (TANF) or a means-tested tax credit such as the Earned Income Tax Credit (EITC). For concreteness, we'll talk about a cash increase but we could just as easily consider a cash decrease; the MVPF would be the same number because both willingness to pay and cost would be negative. For this initial example, we make several assumptions that we'll relax later in the discussion: a cash (not in-kind) transfer; the policy change is small; individuals exhibit privately optimal behavioral responses to the policy change; and no impacts of the policy on the people who were not the policy's direct recipients.

Costs

Consider first the denominator of the expression for the MVPF: What is the cost of increasing the program's cash benefits by $1? It is useful to think of two different classes of cost: the mechanical cost and the fiscal externality. The "mechanical cost" of the policy is the increase in government expenditures due to the policy in the absence of any behavioral response. If the number of infra-marginal individuals who were already receiving the cash transfer policy is I, then the mechanical cost of increasing payments by $1 for each infra-marginal recipient will be $\$I$.

The "fiscal externality" from the policy captures the effect of any behavioral response to the policy on the government's net budget outlays. For example, if individuals reduce their labor supply to become eligible for additional welfare benefits, this will reduce tax revenue collected by the government on earnings. Conversely, if individuals enter the labor force in order to become eligible for an expanded Earned Income Tax Credit and this decreases their use of other government

transfer programs such as food stamps, this will increase net government revenue (but also raise government costs from increased EITC payments). The fiscal externality must account for the full impact of any behavioral response by both marginal and infra-marginal recipients on government spending and tax revenue: for example, it must consider changes in government net revenue arising from changes in eligibility for (and hence spending on) other public programs, changes in sales taxes from modified consumption patterns, changes in public health care spending through Medicare and Medicaid (if the program affects health), and so on.

This concept of a policy's fiscal externality is where the applied econometrics literature on causal inference connects directly with welfare analysis. The fiscal externality logic clarifies which causal effects are and are not necessary for estimating program costs for purposes of welfare analysis. Specifically, it is sufficient to estimate the net impact of the increase in benefit levels on net government revenue, without decomposing the impact into these various channels (Kleven and Kreiner 2005). For example, a large literature has analyzed a wide variety of potential effects of an increase in the level of unemployment insurance benefits on a range of behaviors including unemployment duration (for a recent review, see Shmieder and von Wachter 2016), exit rates into unemployment (Jäger, Schoefer, and Zweimüller 2019) and re-employment wages (Nekoei and Weber 2017). For welfare analysis, however, one needs the net impact of these behavioral changes on the government budget; the individual channels of response are neither necessary nor sufficient.

Benefits

Now consider the numerator of the expression for the MVPF: the benefits from the $1 increase in cash transfer—that is, the willingness to pay by recipients out of their own income for $1 more of the cash transfer. In many cases, this is harder to estimate than the costs of the program (although, deliberately, it is not hard in our first example). But a key insight of the MVPF framework is that they need not depend directly on behavioral responses to—or causal effects of—the policy.

It is useful to distinguish between two classes of recipients of the transfer. For the infra-marginal recipients who were already receiving the cash transfer, the $1 transfer is valued at $1. How much would you be willing to pay for an extra dollar? One dollar. But for the marginal recipients who change their behavior in response to the change in policy and thus become newly eligible for the transfer, how much do they value their new benefit? For example, suppose they decrease the amount they work to become income-eligible for the cash policy. Well, if they are making privately optimal decisions, they must be indifferent to changing their behavior. Why? Because they had already chosen their behavior (in this example, the hours they work) at the optimal level—balancing private costs and benefits of another hour of work—under the old policy, and we've made only a very small ($1) change in the policy. More generally, if the policy change is small and individuals are making privately optimal decisions, the private cost of whatever behavioral change

the marginal recipient undertakes to become eligible for the benefit is equal to the private benefit from becoming eligible.[2]

Given these assumptions, the $1 increase in the cash transfer has no welfare effect on marginal beneficiaries (in other words, those who change their behavior in response to the policy change in order to become newly eligible). The willingness to pay is just $1 times the number of infra-marginal beneficiaries (I). Note that we have not needed to estimate any causal effects to determine the benefits of the policy for either inframarginal or marginal recipients. In other words, despite a large empirical literature (to which we plead guilty of making contributions) on the potential benefits of public policies—the impacts of unemployment insurance on eventual re-employment wages, the impacts of health insurance on health and consumption smoothing, and so on—in this benchmark case, these studies do not directly inform recipient willingness to pay. Causal effects are needed only for the fiscal externality cost term in the denominator because in that setting, an agent who is making (by assumption) privately optimal behavioral changes in response to the policy change will not internalize the external effects of the policy on the government budget.[3]

Putting it Together

It is convenient to normalize the willingness to pay (the numerator of the MVPF) by the mechanical cost to the government. Recall that this mechanical cost was $\$I$, which was also the willingness to pay for the policy. Thus, the MVPF of a $1 increase in cash benefits is:

$$\text{MVPF}^{\$1} = \frac{1}{(1 + FE)}$$

where FE denotes the fiscal externality of the policy per dollar increase in the mechanical expenditure per infra-marginal beneficiary of the policy. Note that the fiscal externality may be positive or negative; policies may have a positive net effect on the government budget (say, by improving health and reducing public spending on health care) or a negative net effect on the government budget (say, by discouraging work effort).

The MVPF measures a policy's bang-for-the-buck. Every $1 of net spending on a tax cut delivers $MVPF of benefits to the recipients of that tax cut. Conversely, every $1 of net revenue raised through a tax increase imposes a cost equivalent to $MVPF

[2]This is known as the envelope theorem. The envelope theorem guarantees that behavioral responses to marginal policy changes by utility-maximizing individuals do not affect their utility directly; however, when prices do not reflect their resource costs, behavioral responses impose a cost on those bearing the difference between the prices faced by the individual and their resource costs. Behavioral responses to policies therefore have first-order effects on policy *costs*—because of the fiscal externality—but only second-order effects on recipient welfare—because of the envelope theorem.

[3]By the same token, if the behavioral responses to the policy have external effects on other individuals besides recipients of the policy, these effects would also have to be taken into account. We cover this possibility below when we consider cases of "multiple beneficiaries." For now, for simplicity, we assume that government policy is the only pre-existing distortion, and hence the only source of potential "external effects."

to the beneficiaries. Spending more resources on policies with higher MVPFs delivers greater welfare to those beneficiaries per dollar spent; raising revenue from policies with lower MVPFs does so with lower welfare loss to those paying for the revenue.

How to Use the Marginal Value of Public Funds for Welfare Analysis

Now that we know how to construct the MVPF (at least for one specific example; we'll discuss more applications in a moment), what do we do with it? For example, an MVPF of 1.3 means that every $1 of net government spending provides $1.30 of benefits to the beneficiaries of the policy, or in other words, the beneficiaries would be willing to pay up to $1.30 for that $1 policy. But is an MVPF of 1.30 "good"? What about an MVPF of 0.8? In other words, what do we do with the MVPF once we've estimated it?

We start with a special case where welfare analysis with the MVPF is comparatively easy: when a policy's net cost to the government (the denominator) is negative and the willingness to pay (the numerator) is positive. In this case, the government spending "pays for itself." Hendren and Sprung-Keyser (2020) define such policy to have an infinite MVPF.[4] A classic example would be if cutting tax rates increases net tax revenue, perhaps because of an increase in labor supply in response to the lower marginal tax rates. This is often referred to as being on the "wrong side of the Laffer curve" because the government can simultaneously *cut* taxes and *increase* revenue.

However, most government expenditures have net positive costs to the government. In this case, the most straightforward use of the MVPF framework is to compare two policies that seek to transfer benefits to the same group of people. For example, imagine a comparison of two policies designed to transfer resources to lower income individuals: expanding the earned income tax credit (a wage subsidy for low income workers) or expanding cash welfare benefits (a direct cash transfer to low-income individuals). If these policies have the same distributional incidence, then spending more money on the one with the higher MVPF is preferred. For the same cost, it creates more transfers to the targeted group. The higher MVPF policy gets more "bang for the buck." This means that one can construct a budget-neutral policy that increases individuals' welfare by spending more on the policy with the high MVPF, with such policy financed by reduced spending on the policy with the low MVPF.

Of course it is rare that two policies target exactly the same population. We even fudged a bit in our preceding example because potential recipients of cash welfare and of the earned income tax credit are overlapping but not identical groups. If two policies target two different groups, how can researchers use the marginal cost of public funds that they have calculated? One option is to compare the MVPF of the policy to a calibrated MVPF for a modification of the tax schedule

[4] At least one of us has wondered why they define a term that is negative as being infinite. The authors explain that they define it as an infinite MVPF to make clear it's "even better" than any finite MVPF.

at the same region of the income distribution (Hendren 2020). Another option is to use the MVPFs to quantify the tradeoff involved in making a budget-neutral change between the policies; in other words, taking a dollar from one policy and adding it to the other. Given two policies, A and B, spending more money on policy A financed by reduced spending on B generates $MVPF_A$ dollars of welfare gain for policy A beneficiaries and $MVPF_B$ dollars of welfare loss for policy B beneficiaries. So, if $MVPF_A = 1$ and $MVPF_B = 2$, this means one can take \$2 from policy B beneficiaries and generate \$1 for policy A beneficiaries.

Is such a transfer from group B to group A desirable? That depends on how one feels about these two different groups (sometimes referred to as their "social welfare weights"). If one places equal value on \$1 in the hands of A beneficiaries and \$1 in the hands of B beneficiaries, then the transfer from group B to group A would not be desirable—instead it would be desirable to increase spending on policy B and reduce spending on policy A. But if one values giving \$2 to policy-B beneficiaries less than giving \$1 to policy-A beneficiaries, one would prefer spending more on policy A financed by less spending on policy B.

How should one decide whether \$1 to group A is preferable to \$2 to group B? Perhaps by introspection. Or on philosophical grounds (Saez and Stantcheva 2016). We don't have "satisfying" answers because economics don't generally have a comparative advantage at specifying societal preferences. People disagree. But the marginal cost of public funds quantifies the tradeoff, which is a crucial first step in deciding whether one "likes" it or not. And it's where economists can most directly contribute to these interesting and difficult questions.

Relaxing Assumptions

Now that we have some idea of how to construct and use the MVPF, we'd like to walk through a bunch of real-world applications. But before we can do so, we promised that we would discuss how to relax a bunch of the heroic assumptions we made for the sake of our "benchmark" example. Here we go.

What if the Policy Changes Are Large?

We considered a \$1 change to a policy. That was one of two key assumptions needed for the argument that for marginal recipients (that is, recipients who change their behavior in response to the policy change to become newly eligible), we could assume their willingness to pay was zero.

In practice, of course, many policy changes are large, and the approximation that marginal individuals who react to a policy change experience no net benefits may fail, perhaps spectacularly. For large policies, the marginal cost of public funds remains a useful guide, but measuring willingness to pay can be less straightforward because it now requires incorporating some value of benefits to marginal recipients. Kleven (forthcoming) provides a recent discussion of this point and some possible approaches for analyzing large reforms.

Fundamentally, we need an estimate of the marginal recipients' demand curve for the increase in public expenditure: that is, willingness to pay is the area under that demand curve. For a large increase in a public cash transfer, preexisting recipients still value the transfer at its dollar value (a dollar is still worth a dollar), but for recipients who change their behavior in order to access the larger public cash transfer, we need to know their willingness to pay for that cash transfer, net of the utility cost of their behavioral change.

Estimating demand is a bread-and-butter task of empirical economics, so we are in familiar—if sometimes empirically challenging—territory. It is all the more challenging when the good is not typically traded in a well-functioning market, so that demand cannot be directly estimated. In the example above, one standard approach (really just a short cut) is to count 50 percent of the increased transfer payments to marginal recipients; this 50 percent approximation follows from an assumption of linearity in the response function and the geometry of triangles (Plimpton 1800BC and Pythagoras 500BC). This approach is popular for its ease of implementation, if not necessarily, its realism; Finkelstein, Hendren, and Luttmer (2019) and Hendren and Sprung-Keyser (2020) are recent examples. More ambitiously, one can specify and estimate an economic model of behavior and use that to derive the demand system. Below, we discuss an application to the MVPF of an increase in import tariffs, using Fajgelbaum et al.'s (2020) constant-elasticity-of-demand system to estimate the welfare impacts on marginal actors.

What if Behavior Isn't Privately Optimal?

Our assumption that individuals make privately optimal decisions was the second key to being able to ignore welfare consequences for marginal recipients. However, a large literature in behavioral economics suggests that individuals commonly make mistakes. In this case, we can no longer assume that the welfare impact of the policy change for marginal recipients is zero, even if the policy change itself is small. For example, a $1 increase in the cigarette tax may induce people to smoke less; if individuals smoke more than they would like to, their reduction in smoking may provide first-order benefits to them.

Here we find ourselves in the world of behavioral welfare analysis. It's no longer enough to estimate the marginal recipients' demand curve because their choices (demand) may not reveal their preferences. Either the researcher must assert that she "knows" the individual's utility function (for example, Bound et al. 2004; Finkelstein, Hendren, and Luttmer 2019) or try to elicit their true valuations, perhaps by experimentally eliminating bias and then eliciting demand (for example, see Allcott and Taubinsky 2015).

What if the Policy Provides an In-Kind Transfer?

Relaxing the first two assumptions led us to the observation that we needed to consider benefits for marginal recipients. To this point, we have also assumed that the transfers are in cash. This made life easy (or at least, easier) because it seems reasonable to assume that *inframarginal* recipients place a value of $1 on receiving a

cash transfer of $1. However, it is not obvious how inframarginal recipients value a $1 increase of spending on in-kind assistance. In-kind transfers in the form of health care, education, housing, job training, and food are a substantial share of government expenditures in the United States and in other high-income countries (Currie and Gahvari 2008).

Consider an increase in $1 of government spending on an in-kind benefit, such as additional government spending per pupil at public universities. Because we can no longer assume that the mechanical cost is valued by infra-marginal recipients dollar for dollar, we need to estimate the willingness to pay by infra-marginal recipients out of their own income per dollar of the in-kind benefit, *W*. The more general formula for the MVPF is then:

$$\text{MVPF}^{\text{inkind}} = \frac{W}{(1 + FE)}.$$

Here, *W* denotes the willingness to pay per infra-marginal recipient for the increased spending on their education. In the cash case, we knew that $W = 1$; $1 of expenditures in the form of a dollar transfer is valued at $1 by those who didn't change their behavior to receive it. But an in-kind transfer might be valued at less than the expenditure on it (in other words, $W < 1$), if it causes the infra-marginal recipients to consume more of the in-kind good than they would if given cash. Alternatively, an in-kind transfer might be valued at more than the government expenditure on it (in other words, $W > 1$) if the government can provide the good at lower cost than is available on the private market.

Estimating *W* can be relatively straightforward if the transferred good is also traded in the market at observed prices. In that case, estimating the demand curve for the good among the infra-marginal recipients gives us *W*. But inferring *W* becomes considerably more challenging when the expenditure is on an in-kind good or service that is not traded in a market—for example, increases in spending in public school, spending used to reduce pollution, or expanded public health insurance.

Of course, the empirical challenge of estimating willingness to pay when demand is not directly observed is not specific to the MVPF framework. Any form of welfare analysis must grapple with how to estimate the monetized value of specific goods and services. Fortunately, a range of techniques have been productively employed. One is to infer willingness to pay from other market transactions—such as wages if the good is bundled into workplace amenities, or house prices if the good is concentrated locally (for example, Rosen 1974; Greenstone 2017). Another is to calibrate (a fancy word for "make up") a utility function for the goods delivered. This approach has been used, for example, in the literature valuing increased generosity of public unemployment insurance benefits (for example, Gruber 1997) or expansions of public health insurance eligibility (Finkelstein, Hendren, and Luttmer 2019). Another option is for the researcher to ask hypothetical questions to elicit the willingness of individuals to pay for a private good, such as health insurance (for example, Krueger and Kuziemko 2013). Yet another approach is to offer the good at

randomized prices and thus estimate willingness to pay directly, as Fischer, Frölich, and Landmann (2018) did for eliciting the value of health insurance in rural Pakistan. Finally, researchers can estimate the benefits of the policy itself and then attempt to monetize these benefits. For example, improvements in test scores are frequently mapped to monetary values through the relationship between test scores and earnings (as in Kline and Walters 2016) and researchers monetize estimated health benefits by relating their estimates to the value of a statistical life or a quality-adjusted life year (as in Currie and Gruber 1996; Goodman-Bacon forthcoming).

External Effects of Policies

So far we have (implicitly) focused on policies that have effects only on their intended recipients. However, many policies have indirect effects beyond the obvious beneficiaries. For example, health insurance subsidies to low income individuals may reduce uncompensated care costs to hospitals and therefore provide benefits to hospital shareholders as well (Garthwaite, Gross, and Notowidigdo 2018). A tax on carbon may affect not only those who use fossil fuels, but also those who benefit from reduced global warming. Vaccine subsidies may provide benefits not just to those receiving the vaccine, but also to those who do not obtain the vaccine and yet benefit from the reduced spread of a virus.

The MVPF framework readily captures these effects. The key extension is to measure the willingness to pay of everyone in the population affected by the policy, including those indirectly affected by the change in the policy. The formulas remain the same as in the above examples, except that the estimation of willingness to pay for inframarginal individuals now includes people who are not direct recipients of the policy.

For example, consider the MVPF of a $1 subsidy to the price of vaccines, which presumably generates positive (health) externalities on the population. As before, the mechanical cost of the subsidy is simply $1 times the number of infra-marginal recipients (I) who were already receiving the vaccine. The fiscal externality (FE) cost includes any impact of the subsidy on the government budget, for example, through changes in health which may affect other publicly-financed health care expenditures or changes in labor market participation and productivity which may affect income tax revenue.

What about the benefits of this $1 subsidy to vaccines? The group of infra-marginal recipients who were already getting the vaccine value the $1 decrease in its price: $1. Again, the group of marginal recipients who choose to get the vaccine because of the price reduction have no net welfare change because they are indifferent between not receiving the additional subsidy and not changing behavior, or receiving the additional subsidy and changing behavior (under the assumption that they were already behaving in a privately optimal manner). However, the fact that these individuals become vaccinated may generate external effects on the health of the rest of the population. The magnitude of these welfare effects depends on the magnitude and the sum total of any benefits (positive or negative) for the externally affected population from the increased vaccination of marginal recipients,

as measured by the externally affected population's willingness to pay; this is not equal to 1 but rather captures their willingness to pay for the marginal beneficiaries to be subsidized to obtain the vaccine. The more others benefit from the vaccine, the higher the MVPF. One would need to estimate the willingness to pay for non-recipients and calculate the numerator as the average willingness to pay across the infra-marginal recipients (those who value the subsidy at $1) and the externally affected population (those who value the increase in the number of people receiving the vaccine by some amount W that would have to be estimated).

In contrast, a $1 subsidy on carbon emissions could have negative externalities. This means the MVPF of a carbon subsidy will be lower than it would be in the absence of these externalities. In this sense, carbon taxes will impose less welfare loss on individuals per dollar of government revenue raised—it will be a more desirable tax than in the absence of the externalities.

Frequently Asked Questions

Why Doesn't Welfare Analysis Have to Think about How Policies Are Financed and the Distortionary Costs of Taxation (as in the Marginal-Cost-of-Public-Funds Approach)?

As we mentioned near the start of this paper, a common approach to welfare analysis is to try to measure the benefits of a policy change and then compare this to "the cost" of raising revenue to pay for the policy, which in turn is commonly defined as expenditures on the policy multiplied by 1 plus "the" marginal cost of public funds. Conventional wisdom usually places this cost somewhere between 0.3 (Poterba 1996) and a more conservative 0.5 (Heckman et al. 2010). In short, the marginal cost of public funds approach seeks to account for the distortionary cost of raising the tax revenue to finance that expenditure. Presto: welfare analysis.[5]

Most commonly, the marginal cost of public-funds approach imagines that the revenue for an expenditure is raised through a linear tax on income that leads to distortions in behavior. However, it has been recognized that this is not the only way to raise revenue, and as a result, there is no single marginal cost of public funds (Kleven and Kreiner 2006). The marginal cost of funds will vary depending on whether we increase taxes on the rich or reduce transfers to the poor. For some taxes, such as carbon taxes, the marginal cost of public funds is potentially negative because taxing carbon can have large benefits in the long run that offset its costs today.

By contrast, an attraction of the MVPF approach is that it severs spending analysis from revenue-raising analysis. We can then think separately about the MVPF both of the spending policy and of various policies to finance it—including reduced spending on other policies, increases in specific taxes, or deficit financing. Thus, the MVPF approach "closes the budget constraint" by comparing two MVPFs to

[5] Of course, usually some of the costs are just transfers, and those only should be multiplied by 0.3, not 1.3. Presto: "Insightful" public finance seminar comment.

form (hypothetical) budget-neutral policies rather than assuming a specific form of (hypothetical) financing for the policy, as the MVPF does.

Why Don't Researchers Need to Estimate Income and Substitution Effects of the Policy Separately (as in the Marginal-Excess-Burden Approach)?

The other common approach in public finance to welfare analysis is the concept of the deadweight loss of a policy (due to Harberger 1964) and its extension to marginal deadweight loss—also known as marginal excess burden (due to Auerbach 1985; Auerbach and Hines 2002).

The marginal excess burden of a tax change is commonly defined as the welfare impact of conducting the policy and simultaneously requiring that the beneficiaries pay for it through individual-specific, lump-sum transfers (Auerbach and Hines 2002). Because the conceptual experiment involves not only the policy envisioned but also these compensatory transfers, calculating marginal excess burden requires measuring the "compensated" response to the policy that excludes the income effect.

It is well-known that estimation of the marginal deadweight loss can be badly biased if the uncompensated (Marshallian) demand curve is used to measure consumer welfare, rather than the compensated (Hicksian) demand curve (Hausman 1981). As a result, this literature has been steeped in the view that it is essential to separate out income effects from substitute effects of the policy, which is challenging to estimate. Moreover, this approach is based on an unrealistic thought experiment in which individual-specific, lump-sum taxation (a policy instrument that doesn't exist) is used to finance the policy. Once again, the distinction between the MVPF and the marginal excess burden approach comes down to how the budget constraint is closed; here, the marginal excess burden approach imagines hypothetical lump-sum taxes, whereas, as discussed earlier, in the MVPF approach, one compares MVPFs of two policies to form hypothetical budget-neutral policies.

How Does the Marginal Value of Public Funds Framework Account for Policies that Affect a Diverse Group of Beneficiaries?

Policies rarely affect a homogenous group of people. Once there are different kinds of beneficiaries to a policy (either because the direct recipients are a heterogeneous group or because of external effects), welfare analysis needs to take account of the fact that societal preferences over transferring resources to different groups may differ. In terms of the example we discussed above, the beneficiaries of a subsidy for health insurance to low-income consumers may include not only the low-income recipients, but also hospital shareholders.

When a policy affects diverse groups, the MVPF is still constructed as previously described. However, it becomes more difficult to think about whether an MVPF of 0.8 or 1.3 is "good" or "bad." To do so, one wants to take account of societal preferences toward the various recipients *within* the group of beneficiaries of a policy. If (for the sake of concreteness) one places lower social value on providing benefits to hospital shareholders than to low income individuals, then for a given MVPF, the

policy will be less desirable if more of the benefits accrue to shareholders than if they accrue to low-income recipients.[6]

Isn't This an Old Idea That's Been around for a Long Time?

Yes. The core ideas of the MVPF are explored in impenetrable detail in Hendren (2016), which itself notes that the mathematical definition of the MVPF is not new.[7] It was initially proposed by Mayshar (1990), where it was referred to (incorrectly) as the "marginal excess burden." In related work by Slemrod and Yitzhaki (1996, 2001) and Kleven and Kreiner (2006), it is referred to as the "marginal cost of funds" (or "marginal benefit of projects" in the case of expenditures).

Likewise, the idea of a fiscal externality is not new: It traces back at least to Ramsey (1927), although its crystallization and importance has become more salient recently (for example, Feldstein 1999; Saez 2004; Kleven and Kreiner 2005; Chetty and Saez 2010). Finally, the key insight that when small shifts in incentives lead to behavioral shifts, the net welfare effect on individuals is zero has been used extensively in previous empirical welfare analyses, including Harberger (1964). Our desire to clarify and illustrate the approach lies not in its novelty but in its usefulness: the fundamental novelty of the MVPF approach is not its mathematics, but its empirics: it relies on the causal effects of the policy and therefore provides a path to welfare analysis that leverages the tools generated in the credibility revolution.

Applications

In this section, we aim to reinforce the ideas behind the MVPF approach, as well as the usefulness of this approach, by giving some examples of how it has or can be applied in a variety of fields.

Income Tax Rates

A classic question in public finance concerns analysis of changes in marginal income tax rates. The MVPF of a tax cut that targets a particular income group tells you the welfare gain to those beneficiaries per dollar of net cost to the government. The benefits (numerator) of a tax cut are straightforward: cutting taxes by a dollar increases welfare by a \$1 (that is, \$1 is valued at \$1 by individuals who would be in that income group even without the tax cut). This \$1 valuation requires us

[6]Formally, Hendren and Sprung-Keyser (2020) show that one needs to use the incidence-weighted average social welfare weight when comparing MVPFs across policies.

[7]With apparently little sense of irony, Hendren (2013) notes in the working paper version: "Relative to [the existing] literature, the primarily contribution of this paper is a clarification." In turn, the current article is a revised version of a teaching note (Finkelstein 2019) in which the mathematical derivations of the MVPF is attributed to Hendren (2016), because it is apparently a natural tendency to attribute an idea to the source from which one learned it. Finkelstein learned this tendency from Scott Stern—we therefore wish to cite Scott appropriately here.

to assume that individuals are making privately optimal decisions so that we can ignore any benefits to marginal recipients who change their behavior in response to the tax cut.

The cost of the tax cut is the sum of the mechanical cost and the fiscal externality. The mechanical cost of the tax cut—that is, the cost per inframarginal recipient, holding behavior constant—is a dollar. The fiscal externality of the tax cut is how the tax cut affects the government budget. Possible behavioral responses may include changes in labor supply and changes in the use of tax sheltering strategies, among others. The key is the elasticity of taxable income (and hence tax revenue) with respect to the tax rate (Feldstein 1999).

This causal object has been the subject of a vast empirical literature in labor economics and public finance. For example, drawing on existing causal estimates of various tax reforms, Hendren and Sprung-Keyser (2020) estimate that for every $1 of revenue raised from the 1993 tax increase on top earners, the government lost $0.46 in revenue from behavioral responses that reduced top earners' taxable income. Therefore the net "cost" of the tax increase on the government budget (mechanical cost of plus fiscal externality of –$0.46) is $0.54, for an MVPF of 1.85 (= 1/0.54). The parameter for a tax increase can be used in reverse to think about a tax cut: that is, a dollar of tax cuts on high earnings costs less than its mechanical cost of a dollar because increases in labor supply (or decreases in tax shelters), increase taxable earnings and tax revenue. The MVPF of a tax cut on top earners is greater than 1 because $1 in tax cuts generates $1 in benefits but costs less than $1 due to the negative fiscal externality.

Indeed, if the fiscal externality of a tax cut at the top is less than –1, then we are on the "wrong side" of the Laffer curve. Cutting tax rates raises revenue, and the tax cut "pays for itself." Hendren and Sprung-Keyser (2020) calculate that the fiscal externality from the Reagan tax cut of 1981 was –1.51, so that the tax cut "paid for itself"—although they caution that there is a wide degree of statistical uncertainty in the estimates of the behavioral response (for example, they can't statistically reject an MVPF of 1).

By contrast, a tax cut at the bottom of the income distribution—say, in the form of an increase in the Earned Income Tax Credit—has a different fiscal externality. When individuals at the bottom of the income distribution enter the labor market, they impose a negative externality on the government budget by taking EITC benefits (which increase government spending), but a positive benefit by taking less in transfers that would go to those with lower incomes (in the form of welfare, food stamps, and other benefits). On net, the calculations in Hendren and Sprung-Keyser (2020) suggest the reduction in transfer payments slightly outweighs the increased EITC costs, so that a $1 mechanical increase in the EITC leads to a fiscal externality that reduces net government costs by .08. This implies an MVPF of $1.12 (= 1/(1 – .08)).

It is perhaps not surprising that the MVPF appears to be lower for a tax cut to the poor than to the rich: this outcome is what would be expected in an optimal tax system set by a planner that places greater social welfare weight on the

marginal value of resources for the poor than the rich. The "bang for the buck" is higher for tax cuts at the top than the bottom, but tax cuts at the top may not be desirable given the greater social value of resources at lower incomes. It is cheaper to raise revenue from the poor, but this of course has adverse distributional implications.

Education

The government is a large provider and funder of education, especially primary schooling. How do we calculate the MVPF for an increase in school funding? To illustrate this, consider the work of Jackson, Johnson, and Persico (2016), who study the effect of K–12 school spending on children's long-run outcomes. They use variation from school finance equalizations to show that increased spending led to an increase in children's earnings trajectory over their life cycle.

To calculate the MVPF in this context, first consider the costs to the government of the policy. There is the upfront cost from increased school spending. This is offset, however, by any increases in future tax revenue paid by the children as a result of their increased earnings. Hendren and Sprung-Keyser (2020) translate the estimates from Jackson, Johnson, and Persico (2016) into a projection of lifetime tax revenue paid. They find that this increase in tax revenue is actually sufficient to cover the initial spending on education (accounting for real government interest rates of 3 percent—which is of course an assumption), so that the net cost of the policy is negative. This implies an infinite MVPF, regardless of the size of the willingness to pay for the policy; as long as willingness to pay is positive (in other words, the children are personally better off from the additional spending), the policy increases welfare without costing the government any money. As a result, we can skirt the more conceptually and empirically challenging task of estimating the willingness to pay for this increase in education spending; we discussed potential approaches to estimating willingness to pay for in-kind transfers in the extensions section earlier, but are glad not to have to actually implement them here.

More generally, Hendren and Sprung-Keyser (2020) have provided a "library" of estimates of the MVPF for over 100 US expenditure policies, including changes in spending on education, job and vocational training, housing subsidies, food stamps, health insurance, and many more. It would be a useful task to develop a comparable MVPF "library" for public expenditure programs in other countries.

De-Worming

The educational example above assumes the only beneficiaries from the expenditures are the individual students themselves. However, many government programs can have externalities onto others. In the education example, expanded education may increase the earnings of the rest of the population through complementarities in production (leading to a higher MVPF). Conversely, some of the estimated wage gains may come from sorting/signaling and therefore impose negative externalities on others (leading to a lower MVPF).

Here, we provide a specific example of how such externalities are incorporated into the MVPF framework in the context of a health policy implemented in a developing country. This example is due to Baird et al. (2016), who study the impact of school-based de-worming treatments in Kenya. They document that these treatments led to improvements in health and long-run earnings for the children in these schools. In addition, the treatments also provided benefits for students in neighboring primary schools—who did not receiving the de-worming treatment—through reductions in transmissions of infection.

Computing the MVPF in this case would therefore involve measuring the willingness to pay for the treatments not only for the children who were directly treated (and their families/communities) but also people in the neighboring areas who also saw improvements in their health (and school attendance rates). We discussed this earlier when we talked about the possibility that policies may have external effects beyond the direct recipients. However, as with the Jackson, Johnson, and Persico (2016) estimates of spending on K–12 education in the United States, Baird et al. (2016) estimate that the net-cost to the government of de-worming is negative: the long-run tax revenue from increased earnings in adulthood is sufficient to cover the government cost of the de-worming efforts. Once again, we are spared having to calculate the willingness to pay for in-kind transfers. Given the estimated effects, de-worming policy has an infinite MVPF and is a win-win for the government and its citizens. Of course, in other settings where net costs are positive, one would have to estimate the affected individuals' willingness to pay for the de-worming using the methods for estimating willingness to pay for in-kind transfers that we discussed earlier.

Import Tariffs

A classic question in international trade concerns the welfare consequences of import tariffs, such as the 2018 tariffs imposed by the United States on goods from China (Fajgelbaum et al. 2020; Cavallo et al. 2019; Amiti, Redding, and Weinstein 2019a). We consider the MVPF of an increase in import tariffs from the perspective of the home country. We therefore ignore any costs or benefits for other countries —this could, of course, be incorporated.

To begin, suppose that an increase in tariffs does not lead to a domestic price change, and there are no retaliatory responses by foreign governments to their tariffs. In this case, a $1 increase in tariffs leads foreigners to pay $1 more in taxes and imposes no costs on domestic citizens. With no change in domestic prices, the willingness to pay by residents in the home country will be zero, resulting in an MVPF of zero. From the home country's perspective, the tariff would be an effective way of raising revenue—or, equivalently, an import subsidy would be a poor use of government revenue.

More commonly though, tariffs increase domestic prices. Indeed, Fajgelbaum et al. (2020), Cavallo et al. (2019), and Amiti, Redding, and Weinstein (2019a) all find that the 2018 tariffs were passed-through in full to domestic prices; in other words, domestic prices went up by the amount of the tariff. In terms of the "benefits"

to US consumers (that is, their willingness to pay to avoid a price increase), individuals would be willing to pay $1 to avoid an increase in prices of $1. If demand for other goods is not affected by the tariff on imports, the "benefits" of the tariff are $1. (Actually, it's –$1, but the denominator will also be negative so they will cancel. Taxes and subsidies on the same good(s) have the same MVPF.)

Turning now to costs, one can think of the tariff as a tax on imported goods, so there is a mechanical cost proportional to the current expenditure on imported goods. But, there is also a potential negative fiscal externality if the tariff reduces consumption of imported goods; the fiscal externality is the impact of this behavioral response to the tariff on tariff revenue. In this case, the net revenue raised by the policy will be less than the mechanical cost. To calculate the fiscal externality, we need a causal estimate of the elasticity of imports with respect to the tariff.

Amiti, Redding, and Weinstein (in this journal, 2019a) estimate that the total government revenue raised by the 2018 tariffs is $15.6 billion; this includes the sum of a mechanical cost of $32 billion and the negative fiscal externality of –$16.4 billion. If the tariff were thought of as "small," the benefits would simply be equal to the mechanical cost of $32 billion; domestic consumers' willingness to pay for $32 billion in revenue is just the increase in revenue, so the MVPF would be simply $32/15.6 = 2.05. But $32 billions is not small. We are now in the world we discussed above: "what if the policy changes are large?" and must try to estimate willingness to pay for non-marginal policy changes.

The approach that Amiti, Redding, and Weinstein (2019a) take is in the spirit of the famous Harberger (1964) triangle: while the first dollar of the tariff raises revenue proportional to $32 billion, as one raises the tariff further, consumers who choose to consume fewer imported goods are less affected by further increases in the tariff. As a result, the last dollar of the tariff imposes a welfare cost of $15.6 billion in contrast to the initial $32 billion. Assuming consumers substitute away from imported goods in a linear fashion, this implies that half of the reduction in tax revenue due to behavioral responses of $16.4 billion is "valued" by consumers. This implies a willingness to pay to avoid the tariffs of $32 billion – $8.2 billion = $23.8 billion. Putting this together implies an MVPF of 23.8/15.6 = 1.5. Every $1 raised by the government imposes a $1.50 (that is, $23.8 billion/$15.6 billion) negative benefit on US consumers.

A further concern from a domestic perspective is that raising tariffs leads to a change in prices of exported goods through terms-of-trade effects. Fajgelbaum et al. (2019) use a trade model to capture the spillover effects of price changes onto substitutes and complements for each product and conclude that US individuals would have been willing to pay a total of $41.6 billion to avoid the increase in tariffs. This differs from the Amiti, Redding, and Weinstein (2019a) number both because it incorporates spillover effects of price changes and because of various implementation choices. Likewise, they estimate a different impact of the tariff on net revenue of $34.3 billion. This implies an MVPF of 1.2 (that is, $41.6 billion/$34.3 billion), so that every $1 of government revenue raised imposes a $1.20 welfare loss on the domestic population because of increased prices and reduced export demand.

Interestingly, these implied MVPFs of 1.2–1.5 for tariffs are in a range similar to that of raising revenue through the income tax.[8] Of course, remember that we have not considered potential policy responses by other countries in the form of a trade war, which would negatively affect domestic consumers in a way we have not captured (but in principle could).

Government Procurement Policy

A classic question in industrial organization considers the optimal design of government procurement contracts (Laffont and Tirole 1993). Empirical researchers have studied public procurement contracts for highways (Lewis and Bajari 2014), defense (Carril and Duggan 2018), health insurance (Decarlois 2015; Cabral, Geruso, and Mahoney 2018), durable medical equipment (Ji 2019), and other goods.

In this case, the MVPF measures the monetary benefit of a change in procurement contract per dollar increase in public costs. To be concrete, consider an increase in the government payment to private insurers to provide insurance coverage to elderly individuals through the Medicare Advantage program. In the United States, individuals eligible for Medicare—the public health insurance program for elderly and disabled individuals—can choose between the publicly-provided, fee-for-service Traditional Medicare program and obtaining subsidized coverage through their choice of a privately-provided Medicare Advantage insurance plan. About 30 percent of the 44 million Medicare enrollees choose Medicare Advantage. One key design question for the government is how much to subsidize purchases of these private plans. Cabral, Geruso, and Mahoney (2018) have analyzed the impact of these subsidies empirically. We would like to analyze the MVPF of a $1 increase in the subsidy per enrollee.

What is the beneficiaries' willingness to pay for a $1 increase in the subsidy? By now, this should be old hat: inframarginal beneficiaries value the $1 transfer at a $1.[9] Marginal beneficiaries are those who switch from Traditional Medicare to Medicare Advantage in response to the increase in subsidy. Cabral, Geruso, and Mahoney (2018) estimate that every $1 of subsidy increases Medicare Advantage enrollment by about 0.09 percentage points. We employ the logic that the marginal actors were indifferent between not receiving the additional subsidy and

[8] It is important to mention one caveat about this result: Further increases in the tariff rate may not actually increase government revenue. In Amiti, Redding, and Weinstein (2019b), the authors note that increasing the tariff from 10 to 25 percent on $200 billion of Chinese imports would lower government revenue, implying an infinite MVPF, so that lowering the tariff would raise welfare.

[9] As Cabral, Geruso, and Mahoney (2018) emphasize, there are two potential types of infra-marginal beneficiaries: consumers who were in Medicare Advantage and insurers who were selling Medicare Advantage. The extra $1 of subsidy from the government to Medicare Advantage may be split between increases in consumer surplus (in the form of lower prices or higher quality) or higher profits to the firms. They estimate the "pass through" rate is about 54 percent to consumers (virtually all of which comes in the form of lower prices) and 46 percent to firms. How does this distributional analysis affect the MVPF analysis? As discussed in the multiple beneficiaries section above, it does not affect the calculation of the MVPF per se, but rather the interpretation of the result.

not switching, or receiving the additional subsidy and switching, so the net welfare change for these switchers is zero, and the size of the enrollment effect does not directly enter the MVPF estimate. Thus, the benefit of the $1 subsidy per existing enrollee is simply $1.

What are the costs of the dollar increase in the subsidy? In the absence of any behavioral response, the mechanical cost of the policy per existing enrollee would simply be $1 as well. Whether the fiscal externality is negative or positive depends on whether Medicare Advantage saves money so that the 0.09 percentage point increase in enrollment leads to an increase or decrease in costs. Existing estimates suggest that the government ends up paying 3–6 percent more for individuals enrolled in Medicare Advantage than it would have if they had enrolled in Traditional Medicare (Medicare Payment Commission 2018; Curto et al. 2014). Even using the 6 percent number would imply that the MVPF of the increase in the Medicare Advantage subsidy is roughly equal to 1 $(= 1/(1 + .0009 \times .06)$.

Conclusion

The MVPF framework offers a powerful approach to empirical welfare analysis of a change in public expenditures or taxes. The approach focuses on the ratio of affected individuals' own willingness to pay for the policy change to the causal effect of the policy on government's net costs.

A key attraction of this approach is that it allows researchers to incorporate causal estimates of policy changes directly into a welfare analysis. In addition, the MVPF provides an important guide for future empirical work on which behavioral responses matter for welfare. Specifically, empirical economists interested in translating the benefits of the "credibility revolution" into progress on applied welfare analysis should focus their efforts on estimating behavioral responses that have fiscal externalities on the government budget, not on behavioral responses whose costs are (approximately) fully internalized by the responding individuals. The approach seems both more robust and easier to interpret than the traditional methods of welfare analysis, which may require estimating effects of hypothetical policies in which those affected are "compensated" for the change through lump-sum transfers.

Of course, the MVPF approach is no panacea. As we emphasized, estimating the willingness to pay for the policy change can be challenging, especially if the policy involves in-kind transfers (such as subsidized education) or effects on individuals not directly targeted by the policy change. Here, we have described how a variety of arrows in the empirical economists' quiver—including structural modeling, calibration exercises, and quasi-experimental or experimental techniques—may usefully be brought to bear. The core value of the MVPF is that it provides clarity on what objects are needed for welfare analysis. In doing so, it can potentially remove the silos across different fields and place welfare analyses on the same playing field.

Only in rare cases will welfare analysis of real-world public policy be clear-cut and straightforward. But the MVPF framework has the flexibility to be applied in a wide range of situations.

■ *We are grateful to Alan Auerbach, Dave Donaldson, Xavier Jaravel, Amy Kim, Henrik Kleven, Enrico Moretti, Matthew Notowidigdo, Ben Olken, Ben Sprung-Keyser, and Sammy Young for helpful comments; to the JEP Editors (and especially Timothy Taylor) for extensive and helpful comments and edits; and to the numerous students and seminar audiences whose (understandable) questions and confusions prompted us to write this essay.*

References

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105 (8): 2501–38.

Amiti, Mary, Stephen J. Redding, and David E. Weinstein. 2019a. "The Impact of the 2018 Tariffs on Prices and Welfare." *Journal of Economic Perspectives* 33 (4): 187–210.

Amiti, Mary, Stephen J. Redding, and David E. Weinstein. 2019b. "New China Tariffs Increase Costs to U.S. Households." *Liberty Street Economics*, May 23. https://libertystreeteconomics.newyorkfed.org/2019/05/new-china-tariffs-increase-costs-to-us-households.html.

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

Auerbach, Alan J. 1985. "The Theory of Excess Burden and Optimal Taxation." In *Handbook of Public Economics*, Vol. 1, edited by Alan J. Auerbach and Martin Feldstein, 61–127. Amsterdam: Elsevier.

Auerbach, Alan J., and James R. Hines Jr. 2002. "Taxation and Economic Efficiency." In *Handbook of Public Economics*, Vol. 3, edited by Alan J. Auerbach and Martin Feldstein, 1347–1421. Amsterdam: Elsevier.

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel. 2016. "Worms at work: Long-Run Impacts of a Child Health Investment." *The Quarterly Journal of Economics* 131 (4): 1637–80.

Bound, John, Julie Berry Cullen, Austin Nichols, and Lucie Schmidt. 2004. "The Welfare Implications of Increasing Disability Insurance Benefit Generosity." *Journal of Public Economics* 88 (12): 2487–2514.

Cabral, Marika, Michael Geruso, and Neale Mahoney. 2018. "Do Larger Health Insurance Subsidies Benefit Patients or Producers? Evidence from Medicare Advantage." *American Economic Review* 108 (8): 2048–87.

Carril, Rodrigo, and Mark Duggan. 2018. "The Impact of Industry Consolidation on Government Procurement: Evidence from Department of Defense Contracting." NBER Working Paper 25160.

Cavallo, Alberto, Gita Gopinath, Brent Neiman, and Jenny Tang. 2019. "Tariff Passthrough at the Border and at the Store: Evidence from US Trade Policy." NBER Working Paper 26396.

Chetty, Raj, and Emmanuel Saez. 2010. "Optimal Taxation and Social Insurance with Endogenous Private Insurance." *American Economic Journal: Economic Policy* 2 (2): 85–114.

Currie, Janet, and Firouz Gahvari. 2008. "Transfers in Cash and In-Kind: Theory Meets the Data." *Journal of Economic Literature* 46 (2): 333–83.

Currie, Janet, and Jonathan Gruber. 1996. "Saving Babies: The Efficacy and Cost of Recent Changes in the Medicaid Eligibility of Pregnant Women." *Journal of Political Economy* 104 (6): 1263–96.

Curto, Vilsa, Liran Einav, Jonathan Levin, and Jay Bhattacharya. 2014. "Can Health Insurance Competition Work? Evidence from Medicare Advantage." NBER Working Paper 20818.

Decarolis, Francesco. 2015. "Medicare Part D: Are Insurers Gaming the Low Income Subsidy Design?" *American Economic Review* 105 (4): 1547–80.

Fajgelbaum, Pablo D., Pinelopi K. Goldberg, Patrick J. Kennedy, and Amit K. Khandelwal. 2020. "The Return to Protectionism." *Quarterly Journal of Economics* 135 (1): 1–55.

Feldstein, Martin. 1999. "Tax Avoidance and the Deadweight Loss of the Income Tax." *Review of Economics and Statistics* 81 (4): 674–80.

Finkelstein, Amy. 2019. "Welfare Analysis Meets Causal Inference: A Suggested Interpretation of Hendren." https://economics.mit.edu/files/16272. (accessed August 1, 2020).

Finkelstein, Amy, and Robin McKnight. 2008. "What Did Medicare Do? The Initial Impact of Medicare on Mortality and Out of Pocket Spending." *Journal of Public Economics* 92 (7): 1644–68.

Finkelstein, Amy, Nathaniel Hendren, and Erzo F.P. Luttmer. 2019. "The Value of Medicaid: Interpreting Results from the Oregon Health Insurance Experiment." *Journal of Political Economy* 127 (6): 2836–74.

Fischer, Torben, Markus Frölich, and Andreas Landmann. 2018. "Adverse Selection in Low-Income Health Insurance Markets: Evidence from a RCT in Pakistan." IZA Discussion Paper 11751.

Garthwaite, Craig, Tal Gross, and Matthew J. Notowidigdo. 2018. "Hospitals as Insurers of Last Resort." *American Economic Journal: Applied Economics* 10 (1): 1–39.

Goodman-Bacon, Andrew. Forthcoming. "The Long-Run Effects of Childhood Insurance Coverage: Medicaid Implementation, Adult Health and Labor Market Outcomes." *American Economic Review.*

Goolsbee, Austan. 1999. "Evidence on the High-Income Laffer Curve from Six Decades of Tax Reform." *Brookings Papers on Economic Activity* 1999 (2):1–64.

Greenstone, Michael. 2017. "The Continuing Impact of Sherwin Rosen's 'Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.'" *Journal of Political Economy* 125 (6): 1891–1902.

Gruber, Jonathan. 1997. "The Consumption Smoothing Benefits of Unemployment Insurance." *American Economic Review* 87 (1): 192–205.

Harberger, Arnold C. 1964. "The Measurement of Waste." *American Economic Review* 54 (3): 58–76.

Hausman, Jerry A. 1981. "Exact Consumer's Surplus and Deadweight Loss." *American Economic Review* 71 (4): 662–76.

Heckman, James, J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010. "The Rate of Return to the HighScope Perry Preschool Program." *Journal of Public Economics* 94 (1–2): 114–28.

Hendren, Nathaniel. 2013. "The Policy Elasticity." NBER Working Paper 19177.

Hendren, Nathaniel. 2016. "The Policy Elasticity." *Tax Policy and the Economy* 30 (1): 51–89.

Hendren, Nathaniel. 2020. "Measuring Economic Efficiency Using Inverse Optimum Weights." *Journal of Public Economics* 187.

Hendren, Nathan, and Ben Sprung-Keyser. 2020. "A Unified Welfare Analysis of Government Policies." *Quarterly Journal of Economics* 135 (3): 1209–1318.

Hicks, J.R. 1940. "The valuation of social income." *Economica* 7 (26): 105-124.

Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico. 2016. "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms." *Quarterly Journal of Economics* 131 (1): 157–218.

Jäger, Simon, Benjamin Schoefer, and Josef Zweimüller. 2019. "Marginal Jobs and Job Surplus: A Test of the Efficiency of Separations." NBER Working Paper 25492.

Ji, Yunan. 2019. "The Impact of Competitive Bidding in Health Care: The Case of Medicare Durable Medical Equipment." https://scholar.harvard.edu/files/yunan/files/dme.pdf. (accessed July 27, 2020).

Kleven, Henrik Jacobsen. Forthcoming. "Sufficient Statistics Revisited." *Annual Review of Economics.*

Kleven, Henrik Jacobsen, and Claus Thustrup Kreiner. 2005. "Labor Supply Behavior and the Design of Tax and Transfer Policy." *Danish Journal of Economics* 143: 321–58.

Kleven, Henrik Jacobsen, and Claus Thustrup Kreiner. 2006. "The Marginal Cost of Public Funds: Hours of Work versus Labor Force Participation." *Journal of Public Economics* 90 (10–11):1955–73.

Kline, Patrick, and Christopher R. Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *Quarterly Journal of Economics* 131 (4): 1795–1848.

Krueger, Alan B., and Ilyana Kuziemko. 2013. "The Demand for Health Insurance among Uninsured Americans: Results of a Survey Experiment and Implications for Policy." *Journal of Health Economics* 32 (5): 780–93.

Laffont, Jean-Jacques, and Jean Tirole. 1993. *A Theory of Incentives in Procurement and Regulation*. Cambridge: MIT Press.

Lewis, Gregory, and Patrick Bajari. 2014. "Moral Hazard, Incentive Contracts, and Risk: Evidence from Procurement." *The Review of Economic Studies* (81) 3: 1201–28.

Mayshar, Joram. 1990. "On Measures of Excess Burden and Their Applications." *Journal of Public Economics* 43 (3): 263–89.

Medicare Payment Commission (MedPAC). 2018. *Report to Congress: Medicare Payment Policy*. Washington D.C.: MedPAC

Nekoei, Arash, and Andrea Weber. 2017. "Does Extending Unemployment Benefits Improve Job Quality?" *American Economic Review* 107 (2): 527–61.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2): 200–49.

Plimpton. 1800 BC. https://en.wikipedia.org/wiki/Pythagorean_theorem#History.

Poterba, James. 1996. "Government Intervention in the Markets for Education and Health care: How and Why?" In *Individual and Social Responsibility: Child Care, Education, Medical Care, and Long-Term Care in America*, edited by Victor R. Fuchs, 277–308. Chicago: University of Chicago Press.

Pythagoras 500 BC. https://en.wikipedia.org/wiki/Pythagorean_theorem#History.

Ramsey, Frank. 1927. "A Contribution to the Theory of Taxation." *The Economic Journal* 37 (145): 47–61.

Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy* 82 (1): 34–55.

Saez, Emmanuel. 2004. "Reported Incomes and Marginal Tax Rates, 1960–2000: Evidence and Policy Implications." *Tax Policy and the Economy* 18: 117–72.

Saez, Emmanuel, and Stefanie Stantcheva. 2016. "Generalized Social Welfare Weights for Optimal Tax Theory." *American Economic Review* 106 (1): 24–45.

Schmieder, Johannes F., and Till von Wachter. 2016. "The Effects of Unemployment Insurance Benefits: New Evidence and Interpretation." *Annual Review of Economics* 8: 547–81.

Slemrod, Joel, and Shlomo Yitzhaki. 1996. "The Costs of Taxation and the Marginal Efficiency Cost of Funds." *International Monetary Fund Staff Papers* 43 (1):172–98.

Slemrod, Joel, and Shlomo Yitzhaki. 2001. "Integrating Expenditure and Tax Decisions: The Marginal Cost of Funds and the Marginal Benefit of Projects." *National Tax Journal* 54 (2): 189–202.

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 168–194

The Persistent Effects of Initial Labor Market Conditions for Young Adults and Their Sources

Till von Wachter

In each recession, a concern arises that young adult workers coming of age in a depressed labor market may bear lasting scars. During the Great Recession from 2007–09, for example, newspaper articles and policymakers voiced concerns for this "Lost Generation." More recently, young adults who had been planning to enter the labor market after graduation found that their job prospects shifted dramatically between, say, December 2019 and March 2020. Of course, most jobseekers and many employed workers suffer during a slack labor market, but young labor market entrants are particularly vulnerable to adverse labor conditions. Young entrants are at the beginning of a very productive phase for their careers, when earnings growth and gainful job mobility of the typical worker are as high as they will ever be for most individuals. Starting out one's working life during a recession can affect unlucky labor market entrants for many years, in some cases, well into middle age.

This article takes stock of what economists have learned about the persistent effects of entering the labor market in a recession. The first section provides a brief discussion of typical earnings experience early in a career, both on average and as affected by cyclical economic patterns, as a benchmark for what follows. We then turn to a standard econometric framework used to discuss the short- and long-term

■ *Till von Wachter is Professor of Economics, University of California at Los Angeles, Los Angeles, California. He is also Faculty Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts; Research Fellow, Centre for Economic Policy Research, London, United Kingdom; and Research Fellow, Institute for the Study of Labor (IZA), Bonn, Germany. His email address is tvwachter@econ.ucla.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.4.168.

effects of initial labor market conditions. The common approach analyzes the shift in earnings-experience profiles in response to cyclical changes in the unemployment rate in the state of labor market entry. We will discuss the assumptions underlying this approach, how it might be implemented using panel and cross-section data, and how it might be adjusted for selection effects.

Over the last 15 years, an increasing number of studies have analyzed the short- and long-term effects on individuals entering the labor market in a recession, and this article will take stock of the core empirical methods and findings from this literature. On average, individuals entering the labor market in a typical recession (a 4–5 point rise in unemployment rates) experience a reduction in earnings of about 10–15 percent initially—somewhat smaller for college graduates, somewhat larger for high school graduates, and a particularly large reduction for nonwhites. Estimates for college graduates suggest that during recessions, workers tend to start jobs at less prestigious occupations and smaller- and lower-paying firms. For some groups, such as PhD economists and possibly MBA graduates, an initial occupation choice permanently affects career outcomes. An early-career economic shock has the potential to be disruptive beyond strictly economic outcomes, too. An increasing number of studies document that adverse labor market entry has effects on health and other outcomes like marriage, divorce, and women's fertility and can affect socio-economic outcomes, health, and mortality in middle age.

Finally, we turn to potential explanations why young entrants to the labor market may be especially vulnerable, and lessons we can learn for the wider labor market. New labor market entrants have a blank slate in terms of work experience, such that typical concerns regarding selection based on prior job or wage histories complicating most other empirical studies of the career impacts of labor market shocks are not present. Hence, empirical studies reviewed here yield experimental estimates that can be used to make inferences on different models of career environment in a literature where causal evidence is typically hard to come by. To do so, we turn to studies that combine two complementary models of career development and take stock in light of the findings of the literature: skill accumulation (for example, Mincer 1974; Ben-Porath 1967) and job search (for example, Burdett 1978; Manning 2003). One framework is that workers first choose an occupation, then invest in occupation-specific skills, and look for a good job in that occupation. Another framework is that individuals have some general skills, look for a good job, and at the same time learn about and choose an occupation, or at least experience some form of growth of earnings on the job. In either of these cases, an initial shock may disrupt career development.

Studying labor market entrants offers interesting insights for economists beyond career development. In the final section, we discuss potential welfare effects, the role of social insurance, and potential lessons from the study of non-economic outcomes. Studying labor market entrants can also provide insights about which job characteristics and worker mobility changes in recessions, provide credible estimates of hysteresis in the labor market, and yield estimates of the costs of business cycles.

Background on Career Profiles

Most individuals transition from full-time schooling, with only part-time or intermittent employment, to seeking full-time work. During the first ten years in the labor market, on average, wages for young workers rise by about 60 percent and earnings rise by about 100 percent (for example, Card 1999). The difference is accounted for by increases in labor supply. The early career period is also very active in terms of job mobility. On average, individuals have seven employers in the first ten years in the labor market (Topel and Ward 1992), a pattern that has remained remarkably stable over time. For the typical worker, this job mobility leads to better jobs at higher paying employers (Smith and von Wachter 2019). After that, mobility and wage growth slow down considerably and for most workers are steady for the next 20 or so years.

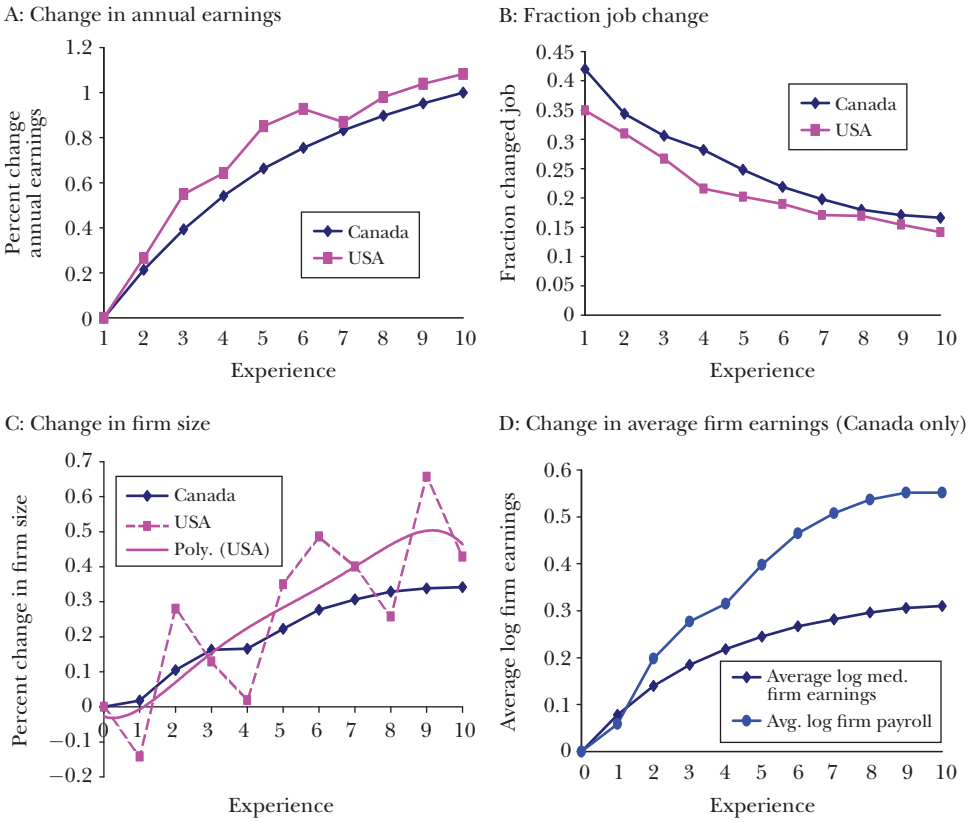
Figure 1 shows some typical patterns for college-educated workers in Canada and the United States from Oreopoulos, von Wachter, and Heisz (2012), though the profiles are similar for other education groups. For each of the four panels, the horizontal axis shows labor market experience, which in this case is just calendar years since graduation. The first two panels show cumulative growth in annual earnings and declining rates of year-to-year job mobility, respectively. The bottom two panels show how two typical measures of firm "quality"—firm size and firm average wages—evolve with labor market experience in Canadian data. Smith and von Wachter (2019) show similar patterns for measures of firm wage for a broader group of US workers.

Figure 2 shows profiles of log annual earnings for annual cohorts of labor market entrants in different years separated by four major education groups from the Current Population Survey. In all years and for all education groups, cohorts' (log) annual earnings evolve in the typical concave fashion seen in Panel A of Figure 1. However, it is also clear that earnings in the initial experience years tend to fluctuate with the business cycle. The vertical lines in each figure show business cycle troughs: notice that the early years of earnings right after such troughs—say, after 2007—are at a lower level. However, this fluctuation becomes more nuanced, or even absent, for workers with higher labor market experience. From the profiles, it is apparent that the initial differences between cohorts tend to fade as cohorts spend more time in the labor market. The empirical strategies and findings discussed later in this paper are effectively based on comparison of these type of profiles of individuals entering in recessions with similar individuals entering in expansions.

At the same time as the rapid school-to-work transition is taking place, most individuals will also have their first experience with other important life events, including cohabitation and marriage, child rearing, and home ownership. For many, this is also a time when social networks of friends and coworkers are formed that can last into middle age and beyond. Theory and evidence (mostly from outside of economics) suggest that the school-to-work transition is related to an important phase of socialization that has lasting

Figure 1

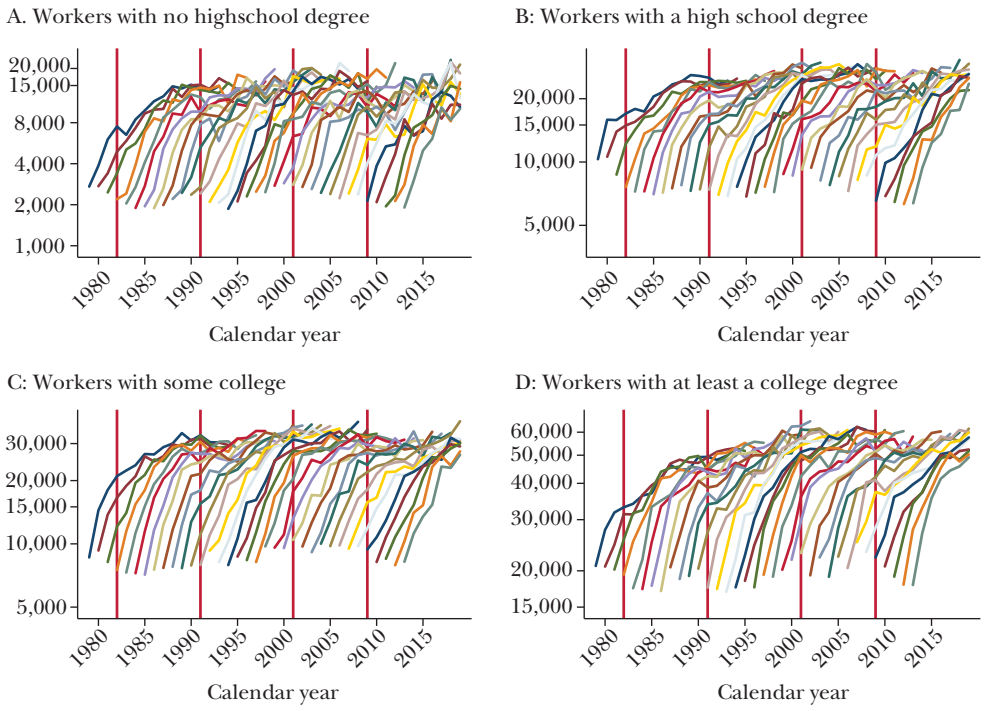
Experience Profiles in Earnings, Mobility, and Firm Characteristics for All Workers with Some College in Canada and the United States



Source: Oreopoulos, von Wachter, and Heisz (2012).

Note: The figure shows average cross-sectional profiles in potential labor market experience (years since graduation) in Canada and the United States; the Canadian figures are derived from the administrative data we use in the paper; the US data are taken from various years of the Current Population Survey (CPS). The underlying sample are all workers with some college in the relevant range of potential experience. Panel A shows percentage increases in annual earnings (for the United States from the March Demographic Supplement of the CPS in 1994–1996). Panel B shows the fraction of workers changing jobs in a given experience year (for the United States, these figures are calculated as the fraction of workers with one year of tenure from the CPS supplements on tenure, mobility, and pensions from 1979 to 2000). Panel C shows the percentage change in firm employment for Canada. This is average firm employment taken over all years the firm was alive from 1982 to 1999, controlling for year fixed effects; for the United States, this is current firm size from firm size brackets taken from Supplements to the CPS in 1979, 1983, and 1988; for the United States, we also show a polynomial approximation). Panel D shows average firm log median earnings or firm log payroll taken over all years the firm was alive from 1982 to 1999, controlling for year fixed effects (see text for details).

Figure 2
Evolution of Average Real Log Annual Earnings for Labor Market Entry Cohorts from 1978 to 2012 by Years since Entry and Major Education Group



Source: Current Population Survey Annual Social and Economic (March) Supplement.
Note: Earnings in US Dollars in 2019 prices, deflated by the Consumer Price Index. Vertical lines drawn at business cycle troughs accordings to https://www.nber.org/cycles.html. Earnings deflated by Consumer Price Index.

influences on attitudes and habits relating to health and substance abuse, among others.[1]

Despite a large amount of descriptive evidence, we still have a limited understanding of the key causal forces underlying the patterns of wage growth shown in Figures 1 and 2. The study of exogenous shocks, such as adverse labor market entry, can help shed additional light on the determinants of career developments, much as the study of mature job losers can shed light on determinants of the wage structure.

[1] As one example, Kaestner and Yarnoff (2011) provide evidence that legal drinking ages in early adulthood can have persistent effects on drinking behavior.

Methods for Estimating Persistent Effects of Initial Labor Market Conditions

The Conceptual Experiment and Baseline Regression

In estimating whether the state of the labor market has persistent effects on earnings, wages, and other outcomes of labor market entrants, the ideal experiment would be to compare two identical groups of entrants that were randomly exposed to different initial conditions. The studies reviewed here seek to approximate this ideal by comparing labor market entrants in different regional labor markets, typically in the United States but increasingly in other countries, that had different unemployment rates.

To understand what empirical variation is effectively used by these studies, it is helpful to consider a typical variant of the regression specification used in this literature. In these regressions, the outcome variable y for individual i would be wages or earnings (often expressed in logs), categorized by the state s and the calendar year t where labor market entry occurred.

$$y_{ist} = \alpha + \gamma_e + \beta_e \, UR_{st} + \theta_t + \lambda_s + \epsilon_{ist}.$$

The main independent variable of interest is the unemployment rate UR_{st} prevailing in a given state during the year of labor market entry. The subscript e stands for years since labor market entry, also referred to as labor market experience. Most studies use so-called potential experience, which is defined as the number of years an individual could have worked after graduation. (In many data sources, it is impossible to calculate actual time worked in any given year.) The basic specification usually includes dummies for potential experience, for calendar year, and the state of labor market entry.

For example, suppose that the outcome variable y is annual labor market earnings. The coefficients on the experience dummies γ_e capture the regular growth in earnings for labor market entrants with experience (the so-called "experience profile" shown earlier). The coefficients β_e on the unemployment rate measure the deviations of earnings from the typical experience profile at each experience year, and hence together capture the shift of the experience profile due to the initial unemployment rate. Given the year dummies and state dummies, the remaining variation in each state's unemployment rate consists of changes over time (relative to its own mean, captured by the coefficient on the state dummy λ_s) that differ from the national business cycle (captured by the coefficients on the year dummies θ_t). These state-specific cyclical changes in the unemployment rate are what identify the shift in the experience profiles due to adverse initial conditions.

Note that this regression equation does not include the state unemployment rate prevailing in the current year t as control variable, but instead includes only the unemployment rate in the year of labor market entry for individual i. In this case, one can show that the coefficient β_e captures the effect of graduating in a recession, given the typical subsequent evolution of local labor market conditions. In

other words, this parameter captures the full difference in lifetime earnings due to adverse labor market entry between lucky and unlucky cohorts.[2]

For many questions, it is important to know what the effect of the very first adverse labor market condition is, net of all subsequent market conditions. It may not be surprising to see extended effects from a career spent in a sluggish economy. Most economic models have a much harder time explaining persistent adverse effects to only short-term exposure to adverse initial conditions. To measure the effect of initial effect net of additional labor market conditions, studies have either included the state unemployment rate in year t in regression model (sometimes referred to as the "current" unemployment rate) or joint dummies for year and state (or whichever labor market area is the focus of the study). Future labor market locations may themselves be affected by initial labor market conditions and possibly correlated with unobservable characteristics, but proceeding at the cell level can help with omitted variable bias.

The regression model is not typically directly derived from a theoretical model. For wages or earnings as an outcome variable, it can be interpreted as a representation of a canonical individual wage or earnings process that incorporates a role of local labor market conditions in a wage setting. Such local effects have been explored empirically and theoretically in the literature on the wage curve, as in Blanchflower and Oswald (1995). Potential economic explanations and models of persistent effects of short-time exposure to adverse initial labor market conditions will be discussed later in this paper.

Several practical choices underlie the decision of this typical regression equation. For example, it uses the unemployment rate in the year of graduation to summarize the state of the local labor market. In case of slow recoveries, alternative measures that are not affected by changes in labor force participation are more suitable, such as the employment-population ratio. The model also presupposes that the effect of the initial unemployment rate is linear. In principle, it is possible that deeper or longer recessions have particularly strong effects. It is also possible that the effect of cyclical conditions changes over time, although analysis of US and Canadian data has revealed that the short- and long-term effects appear comparable in different cycles (for example, Schwandt and von Wachter 2020) and approximately linear in the unemployment rate (for example, Oreopoulos, von Wachter, and Heisz 2012), consistent with similar findings for job losers.

Threats to Internal Validity

A causal interpretation of the coefficient estimates for β_e requires that the economic conditions at labor market entry are uncorrelated with other determinants

[2]The effect of the initial unemployment rate consists of its own direct effect, plus the weighted effect of subsequent unemployment rates correlated with it (for a more detailed discussion, see Oreopoulos et al. 2012). Besides including contemporaneous labor market indicators in the regression, a more complete approach is to allow each subsequent unemployment rate to have persistent effects as well. This is difficult to estimate due to the autocorrelation structure of unemployment rates, but again, it is pursued in Oreopoulos, von Wachter, and Heisz (2012).

of the respective outcome. There are at least two potential threats to validity from estimating this basic regression.

First, individuals may respond to adverse labor market conditions by either anticipating or delaying graduation or by moving to a different local labor market. To address this issue, several papers have used the *predicted* year of graduation—based on age at entry into school or college and duration of the program—to pursue an instrumental variable strategy. In this case, the unemployment rate in the actual year of graduation is instrumented with the unemployment rate in the predicted year of graduation (for example, Kahn 2010; Oreopoulos, von Wachter, and Heisz 2012).

Second, any effect of the initial unemployment rate on wages may also affect labor supply decisions, such that the sample of workers for which the regression can be estimated changes with the business cycle. Most college graduates are likely to work despite adverse labor market conditions, but this can be a more serious issue for less-educated workers. It is possible to assess the extent of the problem by analyzing a sample of individuals that are employed in each year after graduation (as in Oreopoulos, von Wachter, Heisz 2012). In cross-sectional data, a range of strategies can be pursued to address this issue, including imputing small values for zero earnings or wages (so that these workers do not drop out of the sample), quantile regressions, or in the case of earnings, analyzing levels instead of logs such that zeros are included in the regression.

Measurement and Implementation

The most common data source used to study the long-term effects of adverse labor market entry is panel data that records information for the same individuals from the time of labor market entry onwards. Such data have three key requirements: information on the year of labor market entry, the place of labor market entry, and career outcomes for several years after labor market entry. Ideally, the data also have information on the type of education or degree type, so that one can use this information to address potential selection issues (as mentioned in the previous section). Also, it can be useful if the data distinguishes between place of graduation and place of first employment. With these panel data in hand, the regression equation can be estimated directly at the individual level (as in Kahn 2010). However, there are three alternative approaches worth considering.

First, given that the main explanatory variable is defined at the state-year level, it is common econometric practice to work directly at the group level. For example, Oreopoulos, von Wachter, and Heisz (2012) work with means among graduation state, graduation year, and experience cells. If multiple education or demographic groups are studied, the dimension of cells can be changed accordingly. The advantages of the cell-level model are discussed in detail in Angrist and Pischke (2009), among others; in the present context, the ability for graphical representation of the underlying data and results represents a particularly useful feature.[3]

[3] Even if individual-level covariates are used as controls, these can be incorporated in a first step of a regression model. The resulting point estimates are asymptotically equivalent to estimation based on micro data.

A second approach is intended to address a potential drawback of the relatively small sample sizes common with this kind of panel data. While larger administrative panel datasets are becoming increasingly available, they often do not cover enough entering cohorts or have sufficient information on time and place of graduation—and are often not yet universally available to researchers, either. To circumvent these problems, Schwandt and von Wachter (2019) use repeated cross-section data to construct synthetic cohorts of labor market entrants. This approach allows analyzing large cross-sectional data sources previously not available for this type of work, like mortality data from the National Vital Statistics System. A drawback is that the state of labor market entry is typically not observed in cross-sectional labor market or health data. Thus, they have to rely on either the current state of residence or the state of birth to approximate the initial labor market.

To address the potential biases resulting from measurement error and from the possibility of endogenous graduation and mobility between states, Schwandt and von Wachter (2019) use a proxy for the state unemployment rate: they use the state unemployment rate that a birth cohort would have expected to face at labor market entry had it followed typical rates of outmigration from their birth state and typical entry ages (corresponding to four education groups) for their birth state and cohort. They obtain mean migration rates between states and education rates based on state and cohort averages and use them to construct a weighted average unemployment rate that the cohort would have experienced had it followed average migration and schooling patterns. Because this measure does not use actual migration or graduation decisions, it is not affected by endogenous timing or endogenous location of labor market entry.

A third set of data sources has come from the study of specific occupations: for example, MBA graduates, PhD economists, and Japanese college graduates (Oyer 2006, 2008; Kondo 2007). Borgschulte and Martorell (2018) have examined re-entrants after military service, and similar patterns may hold for women returning from maternity leave. This approach can be attractive because certain occupations have well-defined transitions from schooling into the labor market. Occupation-specific studies can also allow a deeper understanding of what mechanism may be driving the persistent effects.

Main Findings on the Persistent Effects of Initial Labor Market Conditions

Result 1: Initial Labor Market Conditions Lower Earnings for 10 to 15 Years

The analysis of college graduates offers a useful benchmark case for the effects of initial labor market conditions, but the qualitative findings also hold for other education groups in the labor market. College graduates typically enter the labor market full time at graduation, making it straightforward to date their expected entry. For most college graduates, it is relatively difficult or costly to postpone labor market entry in the case of recessions. Moreover, college graduates typically

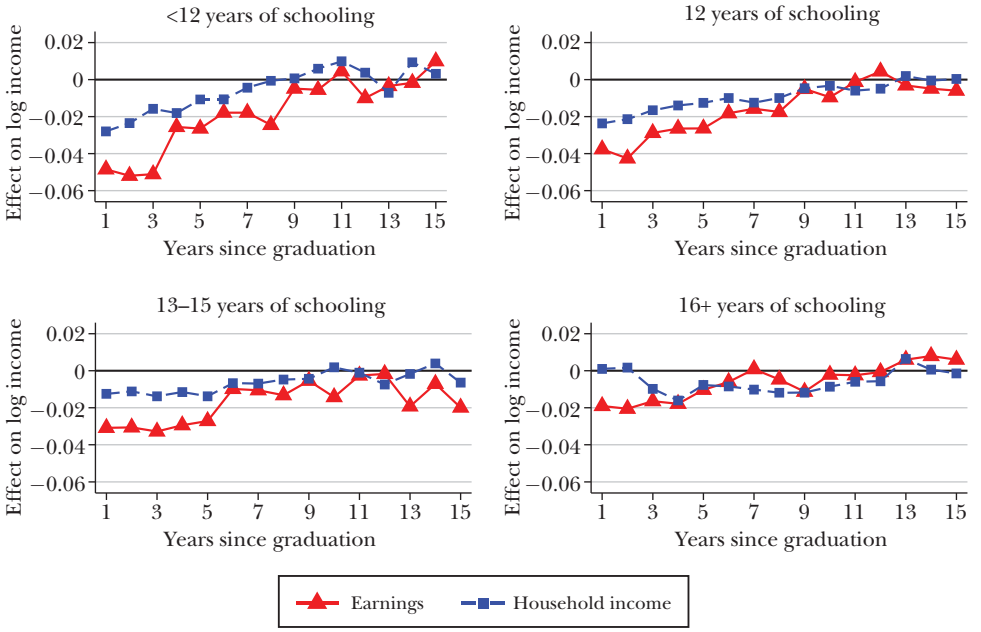
remain in stable employment, so the results are also less likely to be affected by bias from selective labor force participation. These studies suggest that for a typical increase of unemployment in a recession—a rise of 4–5 percentage points in the unemployment rate—the effect of graduating from college in a recession leads to a sharp initial reduction in annual earnings of about 10 percent that fades after about ten years in the labor market (for example, Kahn 2010; Oreopoulos, von Wachter, and Heisz 2012; Altonji, Kahn, and Speer 2016). The initial wage losses seem proportional to the rate of unemployment. Hence deeper recessions, such as the early 1980s recession in the United States (Kahn 2010) or the early 1990s recession in Canada (Oreopoulos, von Wachter, and Heisz 2012), lead to larger initial losses and longer recovery periods that can last up to 15 years. Rothstein (2020) analyzes college graduates specifically entering the labor market during the Great Recession and finds persistent negative effects on earnings and employment. Going beyond college graduates as a benchmark case, Schwandt and von Wachter (2019) analyze the effect of labor market conditions on all workers entering the US labor market from the late 1970s to after the Great Recession and confirm that entering the labor market in a recession leads to persistent effects lasting ten to fifteen years in the labor market.

Result 2: The Size and Duration of Earnings Losses Are Worse for Less Advantaged Workers

When less educated workers start work during a recession, they tend to fare worse initially and experience longer recovery periods as shown in Figure 3, which replicates results from Schwandt and von Wachter (2019). The figure shows point estimates for the change in the experience profile of annual earnings and family income due to a higher state unemployment rate, based on the β_e coefficient obtained from the earlier basic regression equation. The qualitative patterns in the figure are common for studies in this literature. Because labor market entry does not always occur at graduation for less-educated workers, Schwandt and von Wachter (2019) analyze labor market conditions in the predicted year of graduation given the typical duration of education as well as the effect of average labor market conditions at age 18 to 22. With this adjustment, they find that the effect for high-school graduates is about double the effect for college graduates and more persistent. Yet all education groups tend to see a recovery after about ten years in the labor market.

An important question is whether the effect of initial labor market conditions differs by gender, racial, or income groups. Limited sample size makes this a hard question to answer with panel data. Exploiting larger synthetic samples from cross-sectional data, Schwandt and von Wachter (2019) find that non-white labor market entrants experience larger earnings losses, mostly driven by larger reductions in weeks worked in the first four years after labor market entry. However, the persistent effect on hourly wages is of similar magnitude for non-whites and whites. Also, there do not appear to be substantial differences in the effect of adverse labor market entry conditions for women and men.

Figure 3
Effect of State Unemployment Rate at Labor Market Entry on Annual Earnings and Income by Education Groups



Source: Schwandt and von Wachter (2019)
Note: Results are based on the Mincerian specification (equation 2), using data from the Annual Social and Economic Supplement to Current Population Survey from 1976 to 2016.

Result 3: The Effect from Initial Conditions Arises from the Very First Labor Market Condition

Are these earnings losses driven only by the initial exposure or by the ensuing correlated history of unemployment rates? The general finding is that the persistent effects are driven by the very first exposure to unemployment rates alone, though persistent slack tends to lead to longer-term effects. For example, to demonstrate robustness, Oreopoulos, von Wachter, and Heisz (2012) engage in an extensive comparison of the effect of various measures of initial labor market conditions, include state-year fixed effects, and allow for future unemployment rates to have persistent effects. Similarly, Schwandt and von Wachter (2019) show their findings are robust by including the current state unemployment rate as a control variable.

Result 4: The Persistent Earnings Reduction is Largely Driven by Wage Reductions

Are these earnings losses driven by reductions in labor supply, which might imply some form of hysteresis in the labor market, or by persistent declines in hourly wages, which could imply a long-lasting reduction in labor productivity? Schwandt and von Wachter (2019) show that for all education groups, persistent reductions in

wages play a key role in explaining the adverse effect of initial labor market conditions. Employment reductions are less persistent than wage reductions for all workers and generally disappear after about four to five years in the labor market.

Result 5: Unlucky College Graduates Tend to Work in Less Attractive Occupations

Unlucky labor market entrants might end up entering different occupations or otherwise less attractive jobs. In one of the earliest papers in this literature, Okun, Fellner, and Greenspan (1973) suggests that a change in the type of jobs offered over the business cycle may lead to lasting benefits from entering the labor market in a high-pressure labor market. Some papers have indeed pointed to a reduction in jobs in high-wage durable manufacturing sectors during recessions (for example, McLaughlin and Bils 2001), but did not focus on younger workers. Most of the current research on how initial labor market conditions affect occupation and industry has focused on college graduates. For example, Kahn (2010) shows unlucky college graduates start jobs with lower occupational prestige. They also tend to start and stay longer in lower-wage occupations (Altonji, Kahn, and Speer 2016) and industries (Oreopoulos, von Wachter, and Heisz 2012, Web Appendix). Both studies also show that higher-earning majors typically fare substantially better in recessions relative to lower earning majors. Oreopoulos, von Wachter, and Heisz (2012) show that those predicted to have high earnings based on college, major, and degree type fare best in recessions, with only short-term losses, while those at the bottom can experience permanent reductions in earnings. Less information on occupational choice is available for lower educated workers.[4]

Oyer (2006, 2008) study the occupational choices of two groups of high-skilled graduates: Stanford MBAs and PhD economists. In both cases, entering the labor market in a recession has permanent effects on occupational choice. Comparatively unlucky Stanford MBAs are found to have much lower propensities to enter investment banking rather than consulting, and unlucky PhD economists have lower propensities to obtain academic jobs. It is unclear whether such persistence arises because initial investment in job skills specific to an occupation tends to keep a person on a certain career trajectory, or whether perhaps adverse signaling from starting in a less prestigious job hinders unlucky graduates to from switching occupation when the labor market recovers. Nunley et al. (2017) suggest that at least in the short run, signaling from worse quality jobs can play a role for college graduates.

Result 6: Unlucky College Graduates Start Out Working at Less Attractive Firms

Recent work from Haltiwanger et al. (2018) suggests that during a recession, higher-wage firms tend to reduce hiring even within sectors. This may disproportionately affect younger workers, since an important part of their wage growth results from moving to higher paying firms (for example, Smith and von Wachter 2019).

[4] Using cross-country data from European countries, Arellano-Bover (2020a) finds adverse labor market entry leads to persistent reductions in measures of general skills, suggesting a channel along which initial lower job placement could persistently affect wages.

Oreopoulos, von Wachter, and Heisz (2012) exploit their matched worker-firm data to show that college graduates' earnings grow rapidly in part by advancing to higher-paying firms (as shown in Panel C of Figure 1), and that part of the earnings loss from adverse entry arises because the average employer quality is lower for unlucky young graduates. In their data, an above-average rate of job switching after unlucky entry leads to a recovery in firm quality for the first five years, after which graduates' earnings keep recovering while they stay at the same employer. These estimates are replicated in Figure 4, which shows the deviations from the average experience profiles in the outcome variable due to a one-point increase in the provincial unemployment rate. Initial reductions in firm quality are also found in studies of unlucky labor market entrants in Germany (Umkehrer 2019), Austria (Brunner and Kuhn 2014), Norway (Liu, Salvanes, and Sørensen 2016), and Spain (Arellano-Bover 2020b).

Several studies find that mobility between jobs, industries, and occupations tends to be elevated temporarily after adverse entry. The fact that disadvantages from starting at worse initial employers tend to fade suggests that signaling does not appear to be an insurmountable barrier to mobility for the average unlucky graduate. Wozniak (2010), among others, shows that geographic mobility of unlucky college graduates increases and speeds up their recovery, but that the same is not true for lower educated labor market entrants.

Result 7: Persistent Shocks Have Bigger Effects for Entrants than for Mature Workers

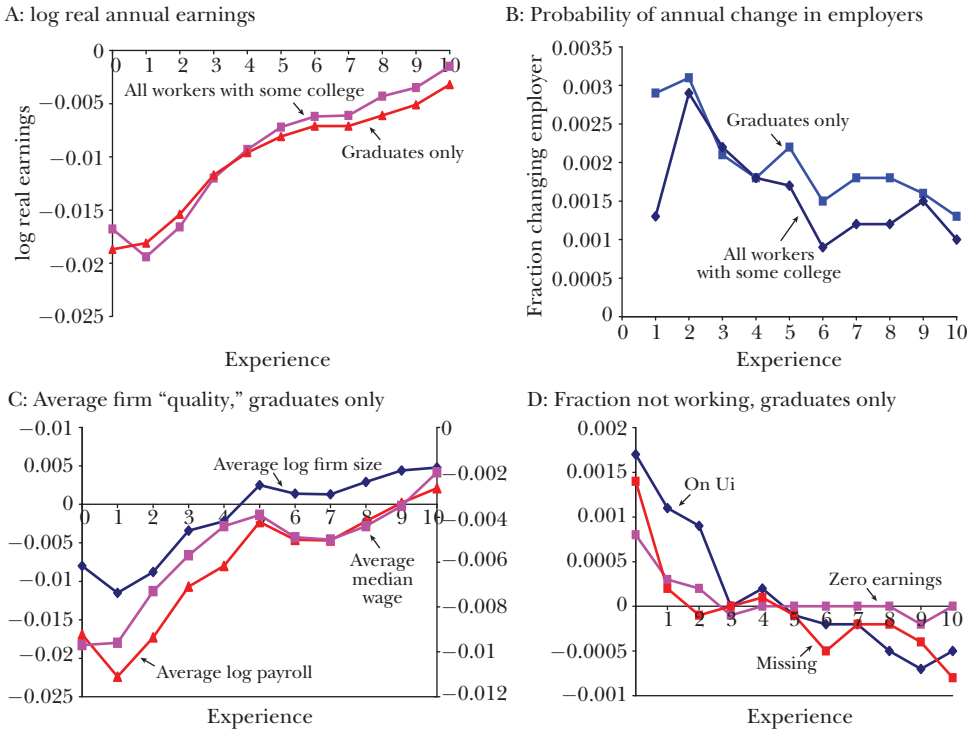
Both aggregate labor market fluctuations and individual shocks can have lasting consequences for more mature workers as well. For the college graduates they analyzed, Oreopoulos, von Wachter, and Heisz (2012) directly compared the effect of unemployment rates at graduation with the persistent effect of unemployment rate at higher years of experience, finding that the initial effect is substantially larger.

This general conclusion is borne out by the related literature. For example, Schwandt and von Wachter (2019) find that each additional point in the initial unemployment rate lowers initial earnings for unlucky labor market entrants by 3.8 percent (see their Table 1). This number can be compared to estimates of the effect of the local unemployment rate on earnings from studies of the wage curve. For example, elasticities for annual earnings reported in Card (1995) imply an approximate marginal effect of 2 percent, about half the effect for labor market entrants. These estimates look only at contemporary effects of labor market conditions on wages. For longer-term effects of the initial unemployment rate in ongoing job spells, Beaudry and DiNardo (1991) find a coefficient of negative 3 percent and Schmieder and von Wachter (2010) report a coefficient of negative 1 percent. In contrast to what Oreopoulis et al. (2012) find for labor market entrants, these papers show that the initial effect gets superseded by subsequent labor market conditions.[5]

[5]The main result of Beaudry and DiNardo (1991) is that on a job spell, the initial unemployment rate eventually gets superseded by the lowest unemployment rate on the job. Schmieder and von Wachter (2010) show this wage premium is lost at job change.

Figure 4

The Persistent Effects of Unemployment in the Year of Graduation on Earnings, Job Mobility, and Firm Outcomes



Source: Oreopoulis, von Wachter, and Heisz (2012).

Note: The figures show coefficients from regressing specified outcome variables on regional unemployment rates at the end of college completion interacted with experience dummies, controlling for effects for cohort of graduation, experience, and region of first residence (equation 4 in the paper). Panels A and B are based on the sample of all 17–20 year olds who started a college program in the data and our main sample of only college graduates. Panel A shows coefficient estimates with log annual earnings as the outcome variable. Panel B shows coefficient estimates using a dummy variable for whether an individual was classified working in a different firm as the one indicated in the previous year as the outcome variable. Panels C and D only show results based on our main sample of college graduates. Panel C shows coefficient estimates using measures of current firm quality as the outcome of interest: the employer's average log total payroll (averaged across all years in the dataset), average log employee size, and average median log wage. Panel D shows coefficient estimates for employment status measures: dummy variables for whether receiving any unemployment insurance in a given year, whether recorded as having zero earnings, or whether not recorded as filing a tax return in a given years. See text for more details.

Studies that consider reemployment wages of older jobseekers also find that the typical effect of poor labor market conditions for more mature workers is smaller than for labor market entrants. For example, Davis and von Wachter (2011) find that each additional point in the unemployment rate lowers percentage earnings losses of workers displaced from stable jobs at mid-size to larger firms by 2.2 percent. Schmieder, von Wachter, and Heining (2019) obtain comparable effects using German data and

show that the majority of the effect is from changes in conditions in the labor market, not changes in worker composition.[6] Thus, wages and earnings of young labor market entrants are substantially more affected by the business cycle than that of more mature job seekers, even though they look for jobs in the same labor market.

Result 8: The Effect of Initial Labor Market Conditions Recur in Middle Age

How long does the effect of initial unemployment rates last? As of the start of 2020, US researchers had studied the short- and longer-run effects of adverse labor market entry for four business cycles back to the early 1980s. To keep a balance between cohorts, most studies focus on the 10–15 year horizon. However, a few studies have attempted to look at longer run effects. For example, Kahn (2010) reports that for college graduates entering during the early 1980s, recession-related earnings losses grow larger again after about 15 years in the labor market. For all recessions since 1970, Rothstein (2020) finds evidence of a lasting effect of the national unemployment rate on unlucky college graduates' long-term employment rate. Schwandt and von Wachter (2020) broaden Kahn's (2010) finding to all unlucky graduates and additional ages and confirm that by the time these unlucky cohorts reach middle age, they have lost ground again in terms of earnings.

Result 9: Poverty of Less-Advantaged, Unlucky Entrants Rises Temporarily Despite More Benefits

The social insurance system is generally ill-equipped to assist unlucky labor market entrants. The typical mechanisms assisting workers in weathering unemployment and earnings reductions, such as unemployment insurance, job search assistance, or retraining, are usually not available to individuals with little labor market experience—because they are likely to have too little employment to qualify for these benefits. Means tested anti-poverty programs, such as Supplemental Nutrition Assistance Program (SNAP) or Medicaid, may provide imperfect insurance for those in most urgent need. In fact, Schwandt and von Wachter (2019) find that receipt of SNAP and Medicaid rises temporarily in response to higher initial unemployment rates, while there is no increase in receipt of unemployment insurance. These increases occur only for workers with a high school degree or less and are substantially higher for non-whites.

Wealth is an important but understudied indicator of the cumulative effect of adverse initial labor market conditions. While Kawaguchi and Kondo (2020) find no effects of adverse labor market conditions on later wealth or homeownership for entrants during the early 1980s recession, descriptive evidence suggests graduates

[6]Using the Displaced Worker Survey, Farber (2011) also shows that reemployment wages of a broader group of displaced workers also fluctuate with the cycle. He does not report point estimates, but from the worst to the best state of the labor market from 1984 to 2010 losses fluctuate from –.2 to –.1. At a 4 to 5 percentage-point difference in the unemployment rate from peak to through, this is in the same ballpark. Estimates for all unemployed job seekers are harder to interpret because of composition changes over the cycle.

entering during the Great Recession may have lower rates of homeownership (for example, Dettling and Hsu 2014).

Result 10: Adverse Early Conditions Worsen Health Behaviors and Raise Long-Run Mortality

Initial labor market entry could affect long-term health for several reasons. One possibility is that certain health behaviors are established in early adulthood; for example, initial labor market conditions persistently increases excessive alcohol consumption (Maclean 2015) and leads to higher obesity and more smoking and drinking in middle age (Cutler, Huang, and Lleras-Muney 2015). More generally, lower income could affect the stock of health through reduced investments and more stress. College graduates entering during the 1980s recession experience higher incidence of heart attacks in middle age (Maclean 2013). Following all labor market entrants from these cohorts, Schwandt and von Wachter (2020) find that starting in their late 30s, unlucky entrants begin experiencing a gap in mortality compared to luckier peers that keeps increasing in their 40s, driven by higher rates of heart disease, liver disease, lung cancer, and drug overdoses.

Result 11: Adverse Labor Market Entry Affects Family Formation, Crime, and Attitudes

Marital patterns of unlucky cohorts are affected from the time they enter the labor market up into middle age, when these cohorts have fewer children (Currie and Schwandt 2014), are more likely to have experienced a divorce, and are more likely to live on their own (Schwandt and von Wachter 2020). Initial labor market conditions also have been found to have effects on attitudes towards economic success and the role of the government (Giuliano and Spilimbergo 2014) and to lead to increasingly lowering individuals' self esteem (Maclean and Hill 2015). Naturally, there is a question of disentangling causality here: lower earnings contribute directly to worsening self-esteem and could affect attitude, but these could in turn help explain lower earnings, something we return to below. Given low incomes and increasing poverty, it is perhaps not surprising that evidence from the United States and United Kingdom indicates that adverse labor market entry persistently raises criminal activities for at least 15 years after entry, especially for men and high-school dropouts (Bell, Bindler, and Machin 2018).

Some Related Studies from Other High-Income Countries

While most empirical work on persistent effects of initial labor market conditions is based on data from North America, the number of studies of initial labor market conditions from other countries is increasing. These studies tend to confirm persistent effects of initial unemployment rates on earnings, employment, and job quality that are greater for lower skilled workers. For example, most college graduates in Japan obtain regular full-time jobs at career fairs at the end of university. If a recession reduces the number of available jobs, and unlucky graduates that do not

obtain such a job the first time around cannot return to the career fair, the effect is a prolonged period of unstable and part time jobs (Kondo 2007). In Germany, most young individuals who are not bound for college enter formal apprenticeship programs that include partial schooling. Unlike US high school graduates who often have a more gradual transition into the labor market, these apprentices have to seek employment once the internship ends, no matter what the state of the labor market. Umkehrer (2019) shows that the initial effect varies by type of training, with manual and service occupation experiencing long-term effects and technical occupations experiencing medium-term effects. In addition, persistent initial effects from adverse initial labor market conditions have also been found in studies from Great Britain (Taylor 2013), Austria (Brunner and Kuhn 2014), Spain (Fernández-Kranz and Rodriguez-Planas 2018), Belgium (Cockx and Ghirelli 2015) and Norway (Raaum and Røed 2006; Liu, Salvanes, and Sørensen 2016; Haaland 2018), among others.[7] In some countries with a rigid wage structure, there are stronger effects on the probability of being employed.

A separate literature has analyzed the long-term "scarring" effects of an early job loss or unemployment spell. Studies based on correlation of initial job instability and longer-term outcomes point to relatively long-lasting effects (for example, Ellwood 1982). However, sustained early unemployment is likely to arise from a combination of exogenous labor market conditions, specific displacement events, and an individual's own characteristics, which can make it hard to obtain causal estimates. Neumark (2002) uses initial local unemployment rates as an instrumental variable and (not surprisingly given the literature discussed here) finds persistent effects of initial job instability. Using year-to-year fluctuations in the retention rate of apprentices in Germany by their training firm as an exogenous displacement shock, von Wachter and Bender (2006) find that initial displacement has a substantial negative earnings effect that fades after about five years. With the increasing availability of large-scale administrative data, additional estimates will likely become available.

Potential Mechanisms Underlying Persistent Career Effects

The empirical results that emerged from the analysis of the effects of initial labor market conditions provide useful experimental findings for the literature on career development that is mostly based on descriptive evidence. These findings can be viewed as empirical moments that any model of career growth has to fit, and in this way, the findings yield binding constraints on existing models. In sum, these trends point to a clear pattern and class of models that future research can refine further.

[7]While all of these papers seek to identify the persistent effects of initial unemployment, they differ in the use of regional versus national unemployment rates, among others. As discussed above, these and other specification choices can matter when comparing magnitudes between studies.

A natural starting point for trying to understand the persistent effects of initial labor market conditions are the two workhorse models of career development: human capital accumulation and job search. It turns out neither of the two can explain persistent effects, but certain combinations of models seem more promising.

Two Benchmark Models of Career Development

The basic human capital model posits that general skills are accumulated on the job, whether through learning-by-doing or on-the-job training (for example, Mincer 1974; Ben-Porath 1967). Such a model could explain depressed wages following longer spells of nonemployment. However, the earlier evidence shows that wages were reduced for up to ten years, even for college graduates who have relatively stable labor force attachment. Also, even short-term exposure to adverse labor market conditions can lead to long-term effects, making it quite unlikely that the basic human capital model could explain these persistent effects.

In a sequential job search model of career transitions, wages of young workers grow as they repeatedly draw job offers from other firms (for example, Burdett 1978; Manning 2003). Given search frictions, a short-term initial reduction in the distribution of wage offers leads to a period of recovery. However, typical estimates of the speed of job offer arrival lead to recovery patterns that only take three to four years (for example, Shimer 2004). This fits with the duration of persistent effects that Oreopoulos, von Wachter, and Heisz (2012) find for their most advantaged college graduates who are perhaps more likely to see only a short-term reduction in job opportunities. But the basic search model has a difficult time explaining more persistent effects of temporary labor market shocks.

Extending the Benchmark Models

As researchers have sought to explain how a short-term initial shock could have longer-term effects, one approach has been to extend the job search model. For Canadian college students graduating during a recession, Oreopoulos, von Wachter, and Heisz (2012) not only observed a lasting decline in earnings over ten years, but also a temporary rise in job mobility and an initial reduction and then recovery in the firm quality of their employers. In particular, increased job mobility and the recovery in firm quality was concentrated among the first five years after labor market entry, suggesting that recovery occurred in two stages: one between firms followed by one within firms. To explain these patterns, they posit a model with two types of labor market entrants (low- and high-skilled) and two types of firms (low- and high-wage). Wages can grow either through finding a job at the high-wage firm or by accumulation of firm-specific skills on the job. Higher-skilled workers have an absolute advantage in job search, and search costs increase with years in the labor market. A labor market shock in this model constitutes a one-period decline in the availability of high-wage jobs. Given the assumptions, a one-period reduction in job availability leads to a persistent effect because workers first search again for a better job. Once workers find a better job, their earnings are lower because they have on

average lower tenure, and wage recovery continues on the job. This model produces persistent losses that eventually fade.

The model captures the reality that switching between firm types plays an important role in explaining career growth (for example, Topel and Ward 1992; Smith and von Wachter 2019), and that the availability of high-wage jobs declines in recessions (for example, Haltiwanger et al. forthcoming). As intended, the model also explains changes in job mobility and job characteristics for unlucky college graduates shown in Figure 4. By design, the model also predicts that low-wage workers take longer to recover, and that they are of higher risk for permanent effects. Remember, search costs in the model rise with time in the labor market. An intuitive reason for this condition might be that as workers marry, have children, and buy homes, costs of job switching rise. Such search costs are less likely to bind for high-wage workers because they have higher job arrival rates. Additional research connecting life events and costs of job search, or on the evolution of job search costs over the life cycle, is needed. But this factor may play a role in the persistent effect of initial labor market conditions.

An alternative and complementary approach has been to extend models of skill accumulation. As one example, to obtain persistent wage effects of initial conditions, Gibbons and Waldman (2006) posit that workers can accumulate general skills and human capital that is task-specific, and that firms create more high-level jobs in economic expansions. In their model, task-specific skills raise wages only in the given job and do not lead to promotion to a higher-level job. They embed these patterns in a model that also has general human capital accumulation, employer learning about worker skills, and comparative advantage. Individuals starting to work in recessions are more likely to start their career in lower-type jobs. While workers may get promoted to higher type jobs based on general human capital accumulation, or be revealed to be of the higher type, once promoted they have less task-specific skills for the higher-type job than luckier labor market entrants who were more likely to start at the higher-type job right away.[8] These effects are smaller for higher educated workers who accumulate skills faster.

Both the Gibbons and Waldman (2006) extension of skill accumulation models and the Oreopoulos et al. (2012) extension of job search models can explain the first four main results of the literature reviewed previously and focus on the role of occupation (Result 5) and firm quality (Result 6), respectively. They can also be used to explain why initially unlucky labor market entrants may experience increasing earnings losses in middle age after an initial earnings recovery (Result 8). In the former case, unlucky entrants have spent less time in the higher job type, and hence have accumulated less task-specific skills. In case of a downturn, these workers would be at higher risk of layoff compared to more lucky entrants. In the latter case, workers

[8] Huckfeldt (2016) also develops a model in which fluctuation of job creation among low- and high-wage occupations over the business cycle can explain why job losers experience larger wage losses in recessions. Neal (1999) analyzes a related model in which workers first search over occupations then look for an employer within that occupation.

have spent less time at their current employer and hence are at higher risk of being laid off again in a downturn. In both cases, a widening of the earnings gap could also arise if in a downturn, lower job- or firm-tenure leads unlucky entrants to experience fewer opportunities for promotion.

While human capital accumulation and search frictions are key ingredients in other models of career growth, some other common ingredients are worth mentioning. One class of models introduces information asymmetries, in which case job mobility can be understood as a process of gradual sorting (as in Farber and Gibbons 1996; Gibbons and Waldman 1999). Another class of models views careers and institutions within firms as an important feature of career development (for example, Doeringer and Piore 1985; Baker, Gibbs, and Holmstrom 1994; Frederiksen, Kahn, and Lange 2020). Research from particular occupations or other countries has pointed to the potential usefulness of considering explanations outside the main economic paradigms.

Broader Implications: Welfare, Policy, and Non-Economic Mechanisms

While research on how outcomes such as health, family status, attitudes, and criminal behavior are affected by unlucky timing of starting a career is still in its infancy, it has begun to yield a richer and more complex picture of the prospects for Lost Generations. Increasing data availability will further improve our understanding of the effect of initial conditions on welfare, the potential for government interventions, and the additional mechanisms explaining the persistent fallout from short-lived economic conditions.

Welfare Effects and Government Intervention

A standard benchmark to gauge the order of magnitude of welfare effects that can be calculated from some of the existing studies is the cumulative loss in short- to medium-term earnings. If individuals make optimal choices conditional on relative prices and their resources, this shift in the budget constraint due to initial conditions will be the key input in a welfare calculation. Table 1 shows measures of the present-discounted loss in earnings over the first 10 and 15 years in the labor market after entry, based on estimates in Schwandt and von Wachter (2019) who present comparable estimates by education and demographic groups based on the same methodology and data. The estimates assume a 5 percentage point increase in unemployment rates, corresponding to a large downturn such as in the early 1980s or the 2008 recession. As a benchmark, the table uses the "lucky" cohort of labor market entrants in the boom year 1995 and discounts future earnings at a rate of 5 percent per year. The table presents the cumulative earnings loss as a fraction of the total present-discounted value (PDV) of earnings over the same period. Among all labor market entrants, the cumulative earnings loss from entering the labor market during a large recession over the first ten years in the labor market amounts to a 9 percent reduction in the present discounted value

Table 1

Losses in Present Discounted Value (PDV) of Annual Earnings in First 10 and 15 Years after Entry into Labor Market, Overall, by Education Groups and Demographic Groups

	Potential labor market experience	*PDV of 1995 entry cohort ("lucky")*	*Loss in PDV due to a large recession*	*Loss as fraction of PDV of "lucky" cohort*	*Loss as fraction of avg. earnings of "lucky" cohort*
All labor market entrants	10	$307,085	–$29,364	–0.096	–0.745
	15	$471,114	–$31,546	–0.067	–0.688
By education group:					
Less than high school	10	$164,278	–$20,870	–0.127	–0.993
	15	$236,536	–$20,979	–0.089	–0.927
High school	10	$222,267	–$23,864	–0.107	–0.841
	15	$329,834	–$24,928	–0.076	–0.784
Some college	10	$264,569	–$25,605	–0.097	–0.758
	15	$398,710	–$32,232	–0.081	–0.835
College or more	10	$458,998	–$22,933	–0.050	–0.391
	15	$699,793	–$19,100	–0.027	–0.281
By Demographic Group:					
Women	10	$273,370	–$23,509	–0.086	–0.674
	15	$405,937	–$25,966	–0.064	–0.664
Men	10	$335,699	–$33,840	–0.101	–0.783
	15	$526,664	–$35,903	–0.068	–0.695
Non-Whites	10	$262,344	–$30,343	–0.116	–0.902
	15	$394,081	–$30,322	–0.077	–0.794
Whites	10	$321,573	–$30,860	–0.096	–0.747
	15	$497,077	–$33,969	–0.068	–0.700

Note: Calculations based on Current Population Survey Annual Social and Economic (March) Supplement and estimates from Schwandt and von Wachter (2019). Potential experience is equal to age minus years of schooling minus 6. PDV is an abbreviation for "present discounted value." The PDV earnings is calculated using a 5 percent interest rate. A "large recession" refers to a rise in unemployment rates in 5 points. To capture workers that made the transition into the labor force, PDV and average earnings are calculated based on annual earnings for workers employed at least 25 weeks in the previous year and with at least 20 usual weekly hours. Average earnings refers to unweighted average of mean annual earnings by experience over respective period (10 or 15 experience years). Dollar values are expressed in 2019 prices using the Consumer Price Index.

of earnings over this period. These losses are larger for less educated workers and non-white workers (with about an 13 and 11 percent loss over a 10-year horizon, respectively), and smallest for college graduates (a 5 percent loss over 10 years). Female entrants tend to have a slightly lower reduction in cumulated present discounted earnings than men (8.6 percent versus 10.2 percent loss over 10 years). These losses amount to three-quarters of mean annual earnings over the first ten years for the average labor market entrant. They range from 100 percent of average earnings for lower educated workers to 40 percent for college graduates.

It is well known that earnings may not fully capture welfare effects for various reasons, including taxes and public or private transfers, costly efforts to prevent career or consumption declines, or direct effects on physical or mental health that are not the result of consumption and investment choices. One estimate of willingness-to-pay to avoid recessions that circumvents these issues comes from reenlisting military personnel. Borgschulte and Martorell (2018) find that military personnel at risk of reentering the labor market in recessions are willing to forego 5 to 7.5 percent of earnings to avoid a recession that raises the unemployment rate by 5 percentage points. Their estimates suggest that individuals bear the cost of over two-thirds of the total present-discounted value of earnings losses from initial labor market conditions, and only one-third is offset by transfers or utility from leisure.

Society's short- and long-run costs from Lost Generations is likely to be larger than welfare losses based on willingness to pay or earnings losses because these measures are unlikely to factor in the full public cost of criminal activity, worsening health, single parenthood, and transfer payments. Only few studies have estimated the effect of early labor market conditions on family income and public and private transfers.[9] Increasing availability of large administrative data sources that integrate information on earnings, family background, social programs, and even taxation will allow a better understanding of changes in income, wealth, and other life circumstances. In addition, these data will allow a better understanding of the role of the tax and transfer system in buffering the effect of initial labor market conditions (for example, Meyer and Wu 2018). Similar integrated data from the criminal justice system may lead to a better accounting of the costs of elevated criminal activity. Such estimates would help to assess whether these young workers would benefit from interventions outside the standard tax and transfer system. This point has been made in the literature of the school-to-work transitions largely outside economics that stresses the role of education in buffering labor market instability (for example, Ryan 2001). Because educational interventions as means of income support have become ubiquitous as a research topic in economics, assessing the potential effectiveness of such interventions for the lower educated, often non-white individuals particularly hard-hit by initial recessions seems worthwhile. A fuller understanding of the economic and health effects well into middle age, for example based on administrative health records, may tilt the balance in favor of such educational interventions.

Additional Mechanisms: Family Formation, Networks, Attitudes, Health

Studies of non-economic outcomes can help refine our understanding of the interaction of economic opportunity, individual choices, and lifetime outcomes. For example, the findings that initial conditions anticipate marriage and fertility, and

[9]For example, using cross-sectional household datasets, Cribb, Hood, and Joyce (2017) find both private and public transfers offset the earnings penalty of initial conditions in the United Kingdom, while Schwandt and von Wachter (2019, 2020) find that increases in public transfers in the United States cannot prevent increases in poverty for less advantaged entrants.

increase divorce and single parenthood, may be a source of increases in the cost of job search that forestalls an earnings recovery through job switching. Similarly, assortative mating within cohorts would lead young workers to have spouses with fewer economic opportunities, further lowering family income and reducing intra-family insurance against shocks. For the same reason, the networks of unlucky individuals' spouses and friends likely have above-average shares of unlucky entrants, potentially affecting economic opportunities (as in Schmutte 2015). At the same time, family networks may play an outsized role as an insurance mechanism for lower-educated unlucky graduates, potentially lowering economic mobility. For unlucky non-white labor market entrants, aggravating factors such as discrimination and incarceration may represent important hurdles to recovery.

All the while, lower lifetime resources, worse health behaviors, and greater stress could make unlucky cohorts respond more strongly to common health shocks whose incidence typically rises with age. Unlucky graduates also highlight how non-economic behaviors could affect economic outcomes. For example, the finding of worsening health behaviors (like excess alcohol consumption, as discussed in Maclean 2015) could explain worsening health (like elevated deaths due to liver and lung disease discussed in Schwandt and von Wachter 2020) and a decline in economic outcomes in middle age. Similarly, changes in attitudes or loss of self-esteem may affect job search or human capital investments. Additional research on these and other potential mechanisms using new data sources would be fruitful and likely help to refine and extend the economic models of career development and economic outcomes we discussed.

Conclusion

Unlucky young workers entering the labor market in recessions suffer a range of medium- to long-term consequences. Large initial effects on earnings, labor supply, and wages tend to fade after 10–15 years in the labor market, partly accompanied by changes in occupation, job mobility, and employer characteristics. Adverse initial labor market entry also has persistent effects on a range of social outcomes, including fertility, marriage and divorce, criminal activities, attitudes, and risky alcohol consumption. Some evidence suggests that early exposure to a depressed labor market lowers health and raises mortality in middle age, patterns accompanied by a reopening of earnings gaps. Overall, the average unlucky college graduate loses about 10 percent of cumulated discounted annual earnings over the first ten years of their career, amounting to three quarters of average earnings during that period. These effects are larger for unlucky lower educated and non-white entrants, who lose up to 13 percent of cumulated discounted earnings and smaller for unlucky college graduates who lose about 5 percent.

Experimental estimates from the analysis of initial conditions on long-term career outcomes can be used to infer about models of career developments. Standard career models fail to explain persistent career effects from short-lived labor

market conditions. Two models combining fluctuations in firm and job quality—sequential job mobility and human capital accumulation—can explain these findings. Additional evidence, possibly from large-scale administrative datasets with detailed information on employers and job characteristics, can be used to refine these models further and to test their predictions against additional career events. The imputation methods proposed by Schwandt and von Wachter (2019) should aid the broader use of datasets that do not have information on place and time of labor market entry. Increasingly available datasets integrating earnings, income, and taxes and transfers can be better used to understand the role of social insurance in preventing hardship among less advantaged labor market entrants and to assess the effectiveness of alternative government interventions. Finally, the increasing research on non-economic outcomes may yield a more integrated understanding of how family formation, social environments, attitudes, and economic opportunities may interact to shape lifecycle outcomes including earnings and health.

The crisis in the labor market triggered by the COVID-19 pandemic has given this line of research increased urgency and has made it relevant to the 4 million or so young individuals graduating from college or high school in the summer of 2020. Some useful lessons emerge from the research reviewed here:

1) Your first job out of school may not be what you had expected, but that's OK. Being flexible in your choice of, say, occupation or where you live will give you more options.

2) Your career will take longer to develop than that of luckier peers. Do what you can to avoid being locked into that first job by continuing to accumulate general skills and looking for opportunities to move to other jobs.

3) If things are going slow, remember, it is hard for everyone. At the same time, all findings discussed here are for averages and do not necessarily apply to you—you have agency in shaping your life and career.

4) You may need to save a higher percentage of income early in life to meet long-term wealth goals.

5) Your desired patterns of marriage and fertility may take more effort to achieve.

6) Take particular care to develop and maintain a healthy lifestyle and be kind to yourself, in part because it will help you weather difficult initial labor market conditions.

■ *I would like to thank Gordon Hanson, Enrico Moretti, Heidi Williams, and Timothy Taylor for very useful comments on a previous draft of the paper and TJ Hedin for helpful research assistance.*

References

Altonji, Joseph G., Lisa B. Kahn, and Jamin D. Speer. 2016. "Cashier or Consultant? Entry Labor Market Conditions, Field of Study, and Career Success." *Journal of Labor Economics* 34 (S1): S361–401.

Angrist, Joshua, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arellano-Bover, Jaime. 2020a. "The Effect of Labor Market Conditions at Entry on Workers' Long-Term Skills." IZA Discussion Paper 13129.

Arellano-Bover, Jaime. 2020b. "Career Consequences of Firm Heterogeneity for Young Workers: First Job and Firm Size." IZA Discussion Paper 12969.

Baker, George, Michael Gibbs, and Bengt Holmstrom. 1994. "The Wage Policy of a Firm." *The Quarterly Journal of Economics* 109 (4): 921–55.

Beaudry, Paul, and John DiNardo. 1991. "The Effect of Implicit Contracts on the Movement of Wages over the Business Cycle: Evidence from Micro Data." *Journal of Political Economy* 99 (4): 665–88.

Bell, Brian, Anna Bindler, and Stephen Machin. 2018. "Crime Scars: Recessions and the Making of Career Criminals." *The Review of Economics and Statistics* 100 (3): 392–404.

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

Blanchflower, David, and Andrew J. Oswald. 1995. *The Wage Curve*. Cambridge, MA: MIT Press.

Borgschulte, Mark, and Paco Martorell. 2018. "Paying to Avoid Recession: Using Reenlistment to Estimate the Cost of Unemployment." *American Economic Journal: Applied Economics* 10 (3): 101–27.

Brunner, Beatrice, and Andreas Kuhn. 2014. "The Impact of Labor Market Entry Conditions on Initial Job Assignment and Wages." *Journal of Population Economics* 27 (3): 705–38.

Burdett, Kenneth. 1978. "A Theory of Employee Search and Quits." *American Economic Review* 68 (1): 212–20.

Card, David. 1995. "The Wage Curve: A Review." *Journal of Economic Literature* 33 (2): 785–99.

Card, David. 1999. "The Causal Effect of Education on Earnings." In *Handbook of Labor Economics*, Vol. 3, edited by Orley C. Ashenfelter and David Card, 1801–63. Amsterdam: Elsevier.

Cockx, Bart, and Corinna Ghirelli. 2015. "Scars of Recessions in a Rigid Labor Market." CESifo Working Paper 5240.

Cribb, Jonathan, Andrew Hood, and Robert Joyce. 2017. "Entering the Labour Market in a Weak Economy: Scarring and Insurance." IFS Working Paper W17/27.

Currie, Janet, and Hannes Schwandt. 2014. "Short- and Long-Term Effects of Unemployment on Fertility." *Proceedings of the National Academy of Sciences* 111 (41): 14734–39.

Cutler, David M., Wei Huang, and Adriana Lleras-Muney. 2015. "When Does Education Matter? The Protective Effect of Education for Cohorts Graduating in Bad Times." *Social Science & Medicine* Elsevier 127: 63–73.

Davis, Steven J., and Till von Wachter. 2011. "Recessions and the Costs of Job Loss." *Brookings Papers on Economic Activity* 43 (2): 1–72.

Dettling, Lisa J., and Joanne W. Hsu. 2014. "The State of Young Adults' Balance Sheets: Evidence from the Survey of Consumer Finances." *Federal Reserve Bank of St. Louis Review* 96 (4): 305–30.

Doeringer, Peter B., and Michael J. Piore. 1985. *Internal Labor Markets and Manpower Analysis*. Armonk, NY: M. E. Sharpe.

Ellwood, David T. 1982. "Teenage Unemployment: Permanent Scars or Temporary Blemishes?" In *The Youth Labor Market Problem: Its Nature, Causes, and Consequences*, edited by Richard B. Freeman and David A. Wise, 349–90. Cambridge, MA: NBER.

Farber, Henry S. 2011. "Job Loss in the Great Recession: Historical Perspective from the Displaced Workers Survey, 1984–2010." NBER Working Paper 17040.

Farber, Henry S., and Robert Gibbons. 1996. "Learning and Wage Dynamics." *The Quarterly Journal of Economics* 111 (4): 1007–47.

Fernández-Kranz, Daniel, and Núria Rodriguez-Planas. 2018. "The Perfect Storm: Graduating in a Recession in a Segmented Labor Market." *Industrial Labor Relations Review* 71 (2): 492–524.

Frederiksen, Anders, Lisa B. Kahn, and Fabian Lange. 2020. "Supervisors and Performance Management Systems." *Journal of Political Economy* 128 (6): 2123–87.

Gibbons, Robert, and Michael Waldman. 1999. "A Theory of Wage and Promotion Dynamics Inside Firms." *Quarterly Journal of Economics* 114 (4): 1321–58.

Gibbons, Robert, and Michael Waldman. 2006. "Enriching a Theory of Wage and Promotion Dynamics inside Firms." *Journal of Labor Economics* 24 (1): 59–107.

Giuliano, Paola, and Antonio Spilimbergo. 2014. "Growing up in a Recession." *The Review of Economic Studies* 81 (2): 787–817.

Haaland, Venke Furre. 2018. "Ability Matters: Effects of Youth Labor-Market Opportunities on Long-Term Labor-Market Outcome." *The Scandinavian Journal of Economics* 120 (3): 794–825.

Haltiwanger, John C., Henry R. Hyatt, Lisa B. Kahn, and Erika McEntarfer. 2018. "Cyclical Job Ladders by Firm Size and Firm Wage." *American Economic Journal: Macroeconomics* 10 (2): 52–85.

Huckfeldt, Christopher. 2016. "Understanding the Scarring Effect of Recessions." https://christopher-huckfeldt.github.io/files/UTSEOR.pdf.

Kaestner, Robert, and Benjamin Yarnoff. 2011. "Long-Term Effects of Minimum Legal Drinking Age Laws on Adult Alcohol Use and Driving Fatalities." *The Journal of Law & Economics* 54 (2): 325–63.

Kahn, Lisa B. 2010. "The Long-Term Labor Market Consequences of Graduating from College in a Bad Economy." *Labour Economics* 17 (2): 303–16.

Kawaguchi, Daiji, and Ayako Kondo. 2020. "The Effects of Graduating from College during a Recession on Living Standards." *Economic Inquiry* 58 (1): 283–93.

Kondo, Ayako. 2007. "Does the First Job Really Matter? State Dependency in Employment Status in Japan." *Journal of the Japanese and International Economies* 21 (3): 379–402.

Liu, Kai, Kjell G. Salvanes, and Erik Ø. Sørensen. 2016. "Good Skills in Bad Times: Cyclical Skill Mismatch and the Long-Term Effects of Graduating in a Recession." *European Economic Review* 84: 3–17.

Maclean, Johanna Catherine. 2013. "The Health Effects of Leaving School in a Bad Economy." *Journal of Health Economics* 32 (5): 951–64.

Maclean, Johanna Catherine. 2015. "The Lasting Effects of Leaving School in an Economic Downturn on Alcohol Use." *ILR Review* 68 (1): 120–52.

Maclean, Johanna Catherine, and Terrence D. Hill. 2015. "Leaving School in an Economic Downturn and Self-Esteem across Early and Middle Adulthood." *Labour Economics* 37: 1–12.

Manning, Alan. 2003. *Monopsony in Motion: Imperfect Competition in Labor Markets.* Cambridge, MA: Princeton University Press.

McLaughlin, Kenneth J., and Mark Bils. 2001. "Interindustry Mobility and the Cyclical Upgrading of Labor." *Journal of Labor Economics* 19 (1): 94–135.

Meyer, Bruce, and Derek Wu. 2018. "The Poverty Reduction of Social Security and Means-Tested Transfers." *ILR Review* 71 (5): 1106–53.

Mincer, Jacob A. 1974. *Schooling, Experience, and Earnings.* Cambridge, MA: NBER.

Neal, Derek. 1999. "The Complexity of Job Mobility among Yound Men." *Journal of Labor Economics* 17 (2): 237–61.

Neumark, David. 2002. "Youth Labor Markets in The United States: Shopping Around vs. Staying Put." *Review of Economics and Statistics* 84 (3): 462–82.

Nunley, John M., Adam Pugh, Nicholas Romero, and R. Alan Seals. 2017. "The Effects of Unemployment and Underemployment on Employment Opportunities: Results from a Correspondence Audit of the Labor Market for College Graduates." *ILR Review* 70 (3): 642–69.

Okun, Arthur M., William Fellner, and Alan Greenspan. 1973. "Upward Mobility in a High-Pressure Economy." *Brookings Papers on Economic Activity* 1973 (1): 207–61.

Oreopoulos, Philip, Till von Wachter, and Andrew Heisz. 2012. "The Short- and Long-Term Career Effects of Graduating in a Recession." *American Economic Journal: Applied Economics* 4 (1): 1–29.

Oyer, Paul. 2006. "Initial Labor Market Conditions and Long-Term Outcomes for Economists." *The Journal of Economic Perspectives* 20 (3): 143–60.

Oyer, Paul. 2008. "The Making of an Investment Banker: Stock Market Shocks, Career Choice, and Lifetime Income." *The Journal of Finance* 63 (6): 2601–28.

Raaum, Oddbjørn, and Knut Røed. 2006. "Do Business Cycle Conditions at the Time of Labor Market Entry Affect Future Employment Prospects?" *The Review of Economics and Statistics* 88 (2): 193–210.

Rothstein, Jesse. 2020. "The Lost Generation? Labor Market Outcomes for Post Great Recession Entrants." Unpublished.

Ryan, Paul. 2001. "The School-to-Work Transition: A Cross-National Perspective." *Journal of Economic Literature* 39 (1): 34–92.

Schmieder, Johannes F., and Till von Wachter. 2010. "Does Wage Persistence Matter for Employment Fluctuations? Evidence from Displaced Workers." *American Economic Journal: Applied Economics* 2 (3): 1–21.

Schmieder, Johannes F., Till von Wachter, and Jörg Heining. 2019. "The Costs of Job Displacement over the Business Cycle and Its Sources: Evidence from Germany." Unpublished.

Schmutte, Ian M. 2015. "Job Referral Networks and the Determination of Earnings in Local Labor Markets." *Journal of Labor Economics* 33 (1): 1–32.

Schwandt, Hannes, and Till von Wachter. 2019. "Unlucky Cohorts: Estimating the Long-Term Effects of Entering the Labor Market in a Recession in Large Cross-Sectional Data Sets." *Journal of Labor Economics* 37 (S1): S161–98.

Schwandt, Hannes, and Till von Wachter. 2020. "Socioeconomic Decline and Death: Midlife Impacts of Graduating in a Recession." NBER Working Paper 26638.

Shimer, Robert. 2004. "Search Intensity." Unpublished.

Smith, Bejamin, and Till von Wachter. 2019. "Job Mobility over the Life Cycle: Evidence from Social Security Records from 1957 to 2014." Unpublished.

Taylor, Mark P. 2013. "The Labour Market Impacts of Leaving Education When Unemployment Is High: Evidence from Britain." ISER Working Paper 2013-12.

Topel, Robert H., and Michael P. Ward. 1992. "Job Mobility and the Careers of Young Men." *The Quarterly Journal of Economics* 107 (2): 439–79.

Umkehrer, Matthias. 2019. "Heterogenous Effects of Entering the Labor Market during a Recession—New Evidence from Germany." *CESifo Economic Studies* 65 (2): 177–203.

von Wachter, Till, and Stefan Bender. 2006. "In the Right Place at the Wrong Time: The Role of Firms and Luck in Young Workers' Careers." *The American Economic Review* 96 (5): 1679–1705.

Wozniak, Abigail. 2010. "Are College Graduates More Responsive to Distant Labor Market Opportunities?" *Journal of Human Resources* 45 (4): 944–70.

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 195–209

Retrospectives
Regulating Banks versus Managing Liquidity: Jeremy Bentham and Henry Thornton in 1802

John Berdell and Thomas Mondschean

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

Great Britain's system of banking and finance was under duress in the 1790s. Revolutionary France declared war with Britain on February 1, 1793, setting in motion a dramatic credit crunch and wave of country bank failures (as detailed in Montefiore 1803). Napoleon Bonaparte had not yet gained political leadership (1799) or become Emperor of France (1804), but as a brigadier general he was leading the French army through victorious battles in Sardinia and what is now northern Italy, and then later on a campaign to Egypt. Meanwhile, Britain was part of coalitions that provided financial and diplomatic support to continental allies fighting against France. Britain's Royal Navy was also involved as in its victories over the French fleet in the Battles of the Nile in 1798 and Trafalgar in 1805.

The Bank of England had become the heart of British finance and had absorbed functions usually associated with a Treasury or Exchequer: for example, it collected taxes, met the government's short-term obligations (Navy and Treasury

■ *John Berdell is Associate Professor of Economics and Thomas Mondschean is Professor of Economics, both at DePaul University, Chicago, Illinois. Their email addresses are jberdell@depaul.edu and tmondsch@depaul.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.4.195.

bills), and paid interest on the public debt. British government borrowing placed tremendous strain on the financial market, and there was a widespread belief that the government prioritized the Bank's contribution to war finance above other monetary policy goals.

In 1797, there was the near-farcical Battle of Fishguard, which remains the most recent episode in which ground troops from Europe invaded Britain. Most of the French army under Napoleon was fighting in central Europe, but the French government emptied some jails and cobbled together 1,400 men whose windblown ship eventually landed them on a beach in Wales. After two days of looting and drunkenness, the invading force surrendered. But when the news of the invasion spread, there was a run by people seeking to convert currency into gold. The British Parliament responded by suspending gold convertibility—the first time Great Britain had done so. As Bordo and White (1991) observe, Britain's financial credibility allowed it to suspend the gold standard while maintaining a strong market for its debt.

While Britain struggled against Revolutionary France, an equally profound revolution was transforming Britain into the world's very first industrial nation. Debate continues to rage over whether military and financial mobilization speeded or slowed industrial transformation, though there is no doubt that it stripped France of her remaining colonies and conclusively demonstrated that Britannia ruled the waves.[1] Contemporaries worried about maintaining financial confidence as Britain's public debt rose (in retrospect to 2.7 times national income) while at the same time industrial expansion relied upon financial innovations that endogenously created money. For example, Jeremy Bentham feared that without bank regulation a "universal bankruptcy" was in the making while Henry Thornton felt that only adroit action by policy makers could avert a "universal failure." In many ways the contrast between their perspectives and recommendations remains with us today, as Bentham urged the regulatory prevention of crisis while Thornton emphasized discretionary policy response.

Henry Thornton was a prominent banker and Member of Parliament. His 1802 essay, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain,* is widely hailed for forcefully arguing that the Bank of England should act as the lender of last resort (for examples, Hicks 1967; Thornton and Hayek [1802] 1939; Woodford 2006). Moreover, Thornton argued that the central bank should take on the responsibility of conducting monetary policy—by which he meant managing to assure sufficient liquidity in the London money market. In contrast, the utilitarian philosopher Jeremy Bentham formulated a remarkably prescient argument for comprehensive bank regulation in an 1801 essay called *The True Alarm*, which

[1] Crowding out of private capital formation by state spending in Great Britain during this time period has been emphasized by an eminent line of historians from Ashton (1959) to Williamson (1984). O'Brien and co-authors (O'Brien and Escosura 1998; O'Brien 2006; O'Brien 2011) find an array of benefits (like demand stimulus and financial development) that may have outweighed costs, while Ventura and Voth (2015) argue that higher interest rates actually spurred industrial transformation by reducing investment in low-productivity agriculture.

included proposals for the establishment and policing of a capital adequacy standard.

Both authors were extremely independent thinkers and we can learn a good deal by reading one against the other. The choice between their approaches continues to generate controversy today. For example, in the aftermath of the Great Recession of 2007–09 when the Federal Reserve and central banks around the world acted as lenders of last resort to major financial institutions, there was an ongoing dispute over whether appropriate banking regulation including higher capital standards could reduce or eliminate the need for such actions in the future. As Swagel (2015) argued in this journal, both international banking standards like the Basel III accords and recent national-level reforms have done much to stabilize the financial system. The existing regulations seek to address both capital and liquidity concerns regarding banks and other systemically important financial institutions with interrelated regulatory ratios on capital, risk coverage, liquidity, stable funding, operational risk, and "total loss absorbing capacity." But despite the shifts in regulations, the coronavirus recession in early 2020 again led central banks to announce plans to act as providers of liquidity and lenders, as well as dealers, of last resort.

In this essay, we start by fleshing out how Britain's financial markets and banking sector were operating in the 1790s, a time of rapidly evolving financial structure repeatedly subject to violent political shocks. We offer a brief overview of Bentham's monetary economics before focusing on his bank regulatory proposal in which he warns of inflation and a looming "universal bankruptcy." We then turn to Thornton's explanation of just how dangerous a collapse of London's money market would be and how liquidity management could avoid it. The final section considers lessons that can be drawn from their contrasting their approaches.[2]

War, Finance, and Innovation

London and the Bank of England formed the center of Britain's financial system, as well as the central node in the international financial network. Bank of England notes were generally restricted to London, where the Bank had a monopoly on their issue. Limited liability companies were rare in Britain, as each needed an act of Parliament. The Bank of England was the only note-issuing English bank that enjoyed a large capital base; other note-issuing English banks were limited to six partners (with unlimited liability), though Scotland had several large banks. As a private bank, the

[2]Works by Bentham, Thornton, and Ricardo will be cited in this essay using compact citations to refer to the appropriate collected works: CW = *The Works of Jeremy Bentham*, published from 1838–1843 under the Superintendence of his Executor, John Bowring; EW = *Jeremy Bentham's Economic Writings, critical edition based on his printed works and unprinted manuscripts*, edited by Werner Stark and published in 1952; PC = Henry Thornton, *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*, with additional writings, edited and with an Introduction by F.A. Hayek, published in 1939; RWC = *The Works and Correspondence of David Ricardo*, Edited by Piero Sraffa with the Collaboration of M. H. Dobb, published in 1951.

Bank of England chose its own leadership and paid (rather high) dividends, lending larger sums to the government each time its charter was renewed. It also paid all short-term government debts (especially Exchequer bills, which would later simply be called Treasury bills) when they came due, regardless of the issuance. A sudden surge here could lead the Bank of England to restrict private credit.

London banks settled transactions in Bank of England notes and looked to the Bank of England as a source of liquidity. "West End" London banks had an aristocratic clientele, while "City Banks" such as Thornton's, were agents for "country" banks. Ashton (1945) distinguishes genuinely rural country banks, which tended to issue locally circulating banknotes, from what were also called country banks in places like Lancashire, the hotbed of the Industrial Revolution. In Lancashire, the circulating medium consisted of bills of exchange, a legally binding promise by one party to make a payment to another party at a certain date, usually no more than a month or two in the future. In Britain, the bill of exchange had undergone a transformation not witnessed in France that rendered it "highly responsive to the community's demand for money, both for transactions and speculative purposes" (Anderson 1970; see also Neal 1994). The bills passed with high velocity from trader to trader, with each party endorsing it to the next, putting their wealth at risk if the borrower failed. Bills commanded greater confidence as the endorsers grew in number and solidity. Bills written for round sums and with solid endorsers circulated with the greatest velocity (Gorton 2020). Merchants with banking connections in London might specialize in discounting bills, which means buying them at less than their terminal (face) value. These banks, or bill dealers, sent some of these bills to London correspondent banks, such as Glyn Mills, for rediscounting. This intermediating tier of London banks held the principal reserves of the country banks and connected borrowers and lenders of loanable funds throughout Britain with the international capital market. The Bank of England routinely rediscounted bills, providing Bank of England notes in their place.

A critical complicating factor here was Britain's anti-usury law, the Statute of Anne, passed in 1713, which limited the nominal interest rate to 5 percent. When the market interest rate rose above 5 percent, the Bank of England became a low-cost source of credit and many packets of bills of exchange would be presented at its discount window. The Bank of England often felt that it needed to ration its discounts, but it might discount freely to allay a panic.[3] Thornton sympathized with the Bank of England's difficult position, noting that a sensitive approach was needed because excessive tightness in discounting would be counterproductive if it induced a panic.[4] James (2012) observes that the United States and the United

[3] Bignon, Flandreau, and Ugolini (2012) suggest that the Bank of England transitioned to free lending and extensive liquidity support against good collateral after 1847 (essentially, following Bagehot's rule before Bagehot wrote it in 1873) while Anson, Bholat, et al. (2017) emphasize the continuation of rationing based on the nature of bills and the parties presenting them.

[4] As we will discuss in the next section, Bentham publicly opposed the law in 1787. So did Thornton, who thought the law increased credit volatility—a view supported by modern writers like Ashton (1959) and Anderson (1970).

Kingdom both had financial systems with a center and a periphery, but the United States lacked a lender of last resort and used illiquid single-name promissory notes. The Bank of England certainly provided a centralized source of liquidity for Britain's market for bills of exchange, but that does not mean that it simply "discounted freely at a penalty rate" as Bagehot (1873) would later describe the actions of a lender of last resort. In any case, James notes that while the United Kingdom experienced many financial crises, its payments system did not episodically freeze up as in the United States.

The growth of bills of exchange in Britain was "exceptionally rapid" in the 1790s (Neal 1994; Pressnell 1956). Moreover, bank owners were often unlimited liability partners in other firms as well as their bank, raising the possibility that a banker's assets might be implicitly pledged to support more than one firm. At this time, for example, the concept of firewalls between separately capitalized subsidiaries of bank holding companies did not exist. The Bank of England's (2020) historical spreadsheet pulls together centuries of macroeconomic and financial data (including that of Broadberry et al. 2015). Using it, we can view the period in which Bentham and Thornton wrote their treatises with modern eyes and the most recently reconstructed data. From 1790 to 1800, real GDP for Great Britain rose by 21.1 percent or at an annual rate of 1.9 percent. The distribution of growth across sectors was varied as agriculture grew only 3.2 percent over the decade, while industry (32.6 percent growth) and services (27.7 percent growth) grew more rapidly. In addition, Broadberry et al. (2015) report implicit price deflators for total GDP as well as for the three sectors. For the 1790–1800 period, the overall price level grew by 48.8 percent or 4.1 percent at an annual rate. Agricultural prices grew the fastest over the decade at 81.7 percent, but industry prices at 34.7 percent and prices for services at 36.7 percent grew as well. Perhaps not surprisingly after the suspension of the gold standard in February 1797, price rises started to accelerate, with the GDP deflator rising by 2.6 percent in 1798, 6.9 percent in 1799, and 11.3 percent in 1800. During these years, the Bank of England expanded its credit to the private sector, while maintaining its holdings of government debt at a relatively constant level (Antipa and Chamley 2017). This is the context during which Bentham's and Thornton's treatises were being written.

Bentham: *The True Alarm*

Bentham's first intervention on economic topics was a public appeal to Adam Smith to reverse his support of the usury laws (discussed in this journal by Persky 2007; also, see Hollander 1999). Bentham argued that the law shifted income from savers to borrowers and lowered capital accumulation. Moreover, he argued that usury laws tended to privilege "old established trades" over "projectors," defined as anyone who seeks to "strike into any channel of invention" or "aim at anything that can be called improvement" (EW I 170, 1787). Bentham subsequently penned a large number of proposals and manuscripts on monetary topics. However, he often

moved quickly from one unfinished work to the next, leaving his devoted editor Étienne Dumont to bring order to the material, much of which remained unpublished for a long time. Bentham implausibly maintained that British Prime Minister William Pitt the Younger only failed to adopt his proposals because Pitt had once acted ungracefully to Bentham after losing a game of chess (EW I 38-9).

In the 1790s, Bentham felt that an increase in paper money could increase the capital stock and output.[5] But by 1801, Bentham had abandoned monetary expansion and was warning that excessive credit creation was leading toward "the greatest of all possible evils, universal bankruptcy: a catastrophe the date of which it is impossible to calculate with precision, but the certitude of which, if no measures are taken to prevent it, can be demonstrated." In comparison to this looming credit crisis, he wrote, "all other questions of political economy have but little importance" (The *True Alarm*, EW III 66). The experience of the 1790s had led him to conclude that inflation invariably accompanied—and was a reliable indicator of—systemic risk.

From a modern perspective, Bentham's diagnosis of the reasons behind the bursts of inflation and financial instability of the 1790s seem off-kilter. But more to his credit, he argued that private bankers do not factor into their behavior what we now refer to as "systemic risk"—that is, the risk that their actions will contribute to an overall financial crisis. As a result, Bentham made the case for what we would now call "microprudential" banking regulation, which refers to the idea that if each bank were required to operate in ways that limit risk for that institution, like holding a capital buffer, then systemic risk would be reduced.

Bentham's diagnosis of inflation and financial instability emphasized behavior by provincial banks rather than action of the Bank of England. He (falsely) believed that only additional banknotes could generate inflation, and thus does not focus directly on the dramatic expansion of bills of exchange in the hands of private bankers in the 1790s. While notes issued by the Bank of England are important components of provincial banks' reserves, Bentham suggests that the provincial banks also held each other's banknotes as reserves. The alarming consequence, he argues, is that the provincial banks are able to expand credit quite independently of the Bank of England. Bentham clearly believes that a reduction in the Bank of England's note issue will not necessarily reduce the volume of provincial notes and lending, and it might even increase it: "If the Bank of England reduced its paper with this motive, would this help to restrain the provincial banks? The abandonment of this profit would be of use only to their rivals . . ." (EW III 189).[6]

[5] See Bentham's *Annuity Notes* plan which Werner Stark dates to 1795–96 (EW II 286). See also Guidi's (2010) account of Bentham's changing assessment of the French Revolution.

[6] Deleplace and Sigot (2011) have shed considerable light on Bentham's approach to banks and money by contrasting it with Ricardo's. When Ricardo was asked by Dumont (Bentham's devoted editor) whether *The True Alarm* should be published, Ricardo opposed publication. For Ricardo, there was no country bank problem: indeed, Ricardo found the autonomy and independence Bentham attributed to country banks quite incomprehensible (RWC III 166-7). In Ricardo's view, the problem lay with the Bank of England and returning to the gold standard. On Ricardo's nuanced support for the gold standard, see Marcuzzo and Rosselli (1994) and Deleplace (2004).

In Bentham's view, "excessive issue" of money is a situation in which the "evils" of inflation and rising bankruptcy risk outweigh the benefits of greater commercial credit. Bentham assumed that provincial banks invariably provided commercial credit (by purchasing bills of exchange at a discounted rate) when they increased their issue of banknotes. The public's propensity to receive and spend the provincial banknotes, rather than return them to the issuing bank, generates what Bentham terms an "air bubble." There is explicit reference to John Law's Mississippi scheme, the world's first international financial bubble that burst in 1720 (EW III 158). According to Bentham, the rate of commercial profit was usually 15 percent while usury laws capped the interest rate banks could charge at 5 percent—and so the demand for paper-creating loans was tremendously strong (EW III 210). Bentham also urged the legislature to begin computing and monitoring the average price level.

Bentham contrasts the "immediate" interest of individual bankers with the "public interest." By making a loan (and issuing more banknotes), a banker increases the risk that the public will experience a catastrophic credit crisis (EW III 189, the editor's uncertain reading of Bentham's handwriting is indicated by [?]):

> Without the intervention of Parliament, individual interest is as favorable to the excess as it can be. Each banker draws [?] his profit in proportion to what he contributes to the excess; in restraining himself he sacrifices all that he could have gained, and what he adds thereby to his own security is almost nothing, as long as he may be engulfed in the catastrophe brought about by the temerity of others.

Bentham appeals to the legislature for regulations, rather than to the executive, because he supposes that Britain's "Minister of Finance" welcomes any increase in output generated by money, even if inflation harms the majority of the population. The balance of political forces does not give Bentham great hope, but it does leave him in no doubt as to where the public interest lies: "*it is evident enough that the loss is for the greater part of the community, and the profit for a very small number*" (EW III 215, emphasis in original).

In setting out his "definitive remedies," Bentham emphasizes the fragility of *confidence*: ill-considered regulatory changes may set off a bank run. For instance, a precipitate banning of small-denomination notes from country banks would lead to a sudden increase in demand for metallic currency; this might have the "appearance" and hence the same consequence as a bank run (EW III 106). Bentham suggests 15 distinct articles, to be enacted in two separate pieces of legislation.

The first act would require the registration of banks and would limit creation of banknotes. Each bank would require a bank charter, which Bentham calls a "patent." He writes: "No patent giving the right to issue bank notes shall be delivered before the banker has furnished security either with or without a pledge, in a certain proportion to the greatest sum of paper money which he is entitled to keep in circulation at any one time" (EW III 175). This legislation would require each

banker to report annually the "average amount of his notes in circulation and the average size of the security fund which he keeps in reserve" (EW III 176). Bentham does not suggest imposing a required reserve ratio; he takes it for granted that fractional reserves are incompatible with what he calls a guarantee of "immediate and uninterrupted solvability." He focuses instead on capital and bankruptcy.

Unfortunately, Bentham's language turns semantically confusing in a key passage. He suddenly mentions a "pecuniary security fund," which might sound like cash reserves but seems intended to mean something more like bank capital (EW III 177-8):

> As for the pecuniary security fund required of each banker, its purpose concerns their final solvability: for we have often repeated that immediate and uninterrupted solvability at every juncture is irreconcilable with the very essence of the banking trade. It is enough to be assured that the value of the pledge is such that the bankers would be able, in the last resort, to meet their engagements. Houses and landed estates which could not contribute to *immediate* solvability may serve as a guarantee for final solvability.

"Immediate" and "final solvability" correspond to the topics that modern economists discuss as liquidity versus solvency. Bankers in the 1790s often owned manufacturing firms, or were involved in trade (EW III 153). Thus, bankers' assets could be pledges as backing for several enterprises. Thus, Bentham wants specific assets identified as bank capital, enough to cover the notes issued by the bank. Bentham insists that the bank capital be reviewed annually (EW III 175). The intention seems to be that in bankruptcy, the pledged assets could be sold, even if slowly. However, one would not today suggest that bank capital consists of an illiquid asset like real estate. Despite these difficulties, Bentham clearly distinguishes "final solvability" from "immediate and uninterrupted solvability." While modern economists have learned that distinguishing liquidity and solvency in the midst of a financial crisis is not easy, Bentham hoped to stave off financial panic and "universal bankruptcy" by creating greater confidence in the "final solvability" of banks in bankruptcy. This approach clearly reflects his lifelong interest in legal reform, and in the corruption and inefficiency of the bankruptcy proceedings of the time (Duffy 1980).

A second essential component of Bentham's proposal is contained in the bank "patents" (or charters). Each will specify "the greatest sum of paper money which he [the banker] is entitled to keep in circulation at any one time." Bentham fixes the number of bank charters and foresees them becoming more valuable over time, with the increasing value of a banking franchise compensating bankers for the regulatory burden they face. "Vacant" bank patents would be auctioned. Bentham clearly recognizes that capital requirements and regulatory supervision will reduce the competitiveness of the market—an outcome recently emphasized by Schliephake and Kirstein (2013). But Bentham had a penchant for compensating a branch of commerce for the imposition of a tax by limiting entry.

Indeed having put an inflexible ceiling on the paper money emitted by the provincial banks, Bentham goes so far as to wonder whether steps might be needed to limit the metallic money supply (EW III 178):

> If, after having put an end to the encrease [sic] of paper money, money in general still continued to multiply, from the augmentation of metallic money, to the point of producing a sensible rise in prices, it would be necessary to take measures to limit the augmentation of metallic money to the degree required for the end in view.

Bentham turned away from financial topics after 1801; consequently, this essay appears to constitute his final position on credit creation. His detailed regulatory program insightfully aims to prevent the buildup of systemic risk due to misaligned incentives. Perhaps if he had the price data he desired, he might not have regarded inflation as an unerring indicator of excessive credit and money creation. But he did hold this view, which led him to the strictest possible monetary rule, that the supply of money and credit should be consistent with the growth of potential GDP and stable prices.

Thornton: *Paper Credit*

Henry Thornton's star has steadily risen since the publication of Hayek's edition of *Paper Credit* in 1939. The first half of his 1802 work emphasizes the need to prevent contractions in the "circulating medium," while in the second half he warns against an excessive increase in Bank of England notes. Chapter VII on "Country Banks—their Advantages and Disadvantages" is crucial to Thornton's transition between these two sections. It is also the most obvious point of comparison to *The True Alarm*. Bentham had tremendous sympathy for Thornton's analysis in *Paper Credit,* declaring to Dumont: "This is a book of real merit." While Bentham quickly saw that they disagreed on the need to control banknotes issued by provincial banks, he felt that "a controversy with him would be really instructive" (EW III 46).

Thornton sets out with a rough count of the number of country banks, documenting their rapid growth in the peaceful interlude between the end of American War in October 1781 and the beginning of the war with France in 1793. Since the onset of war with France, he thinks the number has been stable. Thornton also values the increase in Britain's capital stock that the country banks supported through business loans. Thornton's defense of the Bank of England's management is legendary, but his defense of country bankers is no less impressive. They "take care to lend the sums which have been deposited in their hands, not to the imprudent speculator . . . but to those who . . . manage their concerns with prudence, [and] give proof that they are likely to repay the loan" (PC 175). Indeed the growth of country bankers has made the evaluation of creditworthiness into a "science" that has greatly contributed to British commerce (PC 176). Country bankers have

important informational advantages when they assess credit because "the bill transactions of the neighborhood pass under his view: the knowledge, thus obtained, aids his judgment; and confidence may, therefore, be measured out by him more nearly than by another person" (PC 175).

In the banking system of that time, country banks received loans from London banks, and London banks received loans from the Bank of England. In Thornton's view, the Bank of England's discretion in discounting bills of exchange crowns and stabilizes this prudential hierarchy of credit. Thornton was concerned with the possibility of a "universal failure," but for Thornton this denouement will almost certainly be the result of poor decisions made at the apex of the credit hierarchy rather than at its base.[7] Everything hinges on whether the Bank of England allows the volume of its bank notes outstanding to fall, he argued, because those notes form the means of payment within the London money market. The significance of a liquidity panic in the countryside, such as that which accompanied the outbreak of war in 1793 or the invasion threat of 1797, lies with its effects in London. As trust in country banknotes evaporates, they are replaced by Bank of England banknotes—which are normally confined to circulating in London. "Pressure" builds as the Bank of England notes, which serve as reserves and clearing balances in London, fall relative to large financial transactions.

Thornton suggests that the illiquidity of one important actor could set off a general "alarm" or panic. Because London has become the "general money market" of the country, the danger that a large bank will trigger a cascading payment crisis rises with "pressure": "Some political persons have assumed it to be a principle, that in proportion as the gold of the bank lessens, its paper, or, as is sometimes said, its loans . . . ought to be reduced . . . [a] maxim of this sort . . . would lead to universal failure" (PC 227). Thornton depicted the danger of excessive pressure in London's payments system in alarming terms (PC 114):

> A deficiency of notes in London is a very different thing from a deficiency either of country bank notes or of coin in the country. A large proportion of the London payments are payments of bills accepted by considerable houses, and a failure in the punctuality of any one such payment is deemed an act of insolvency in the party . . . any very great and sudden diminution of Bank of England notes would be attended with the most serious effects both on the metropolis and on the whole kingdom. A reduction of them which may seem moderate to men who have not reflected on this subject—a diminution, for instance, of one-third or two-fifths, might, perhaps, be sufficient to produce a very general insolvency in London, of which the effect would be the

[7] Thornton (PC 210) was on exactly the same wavelength as Ricardo's position discussed in the previous footnote when he asserts: "[I]t has been shewn, that the country paper, however it may fail to be limited in quantity by any moderation or prudence of the issuers, becomes no less effectually limited through the circumstance of their being compelled by the holders to exchange as much of it as is excessive for the London paper which is limited; which is limited, I mean, in consequence of a principle of limitation which the directors of the Bank of England have prescribed to themselves."

> suspension of confidence, the derangement of commerce, and the stagnation of manufactures throughout the country."

Thornton reassures us that usually "steps would be taken" to avert a "general insolvency" of this magnitude: "[T]here is too strong and evident an interest in every quarter to maintain, in some way or other, the regular course of London payments, to make it probable that this scene of confusion should occur; or, even if it should arise, that it should continue" (PC 155). The government's solution to the 1793 panic, for instance, seemed to have worked well: an emergency loan of up to five million pounds of highly liquid bills issued by the Exchequer to "as many mercantile persons, giving proper security, as should apply" restored the "regularity of payment" to the London money market (PC 98-9; for details see Andréadès 1909). This looks very much like an early instance of the Treasury stepping in to supplement, or substitute for, central bank action. Thornton is willing to "hazard" the observation that the Bank should have increased its note issue prior to the "alarm" of 1793 (PC 128). In Andréadès's (1909) telling, the Bank's crude credit rationing rightfully panicked the London market.

Why would the Bank of England allow "pressure" in the money market to develop into "universal failure"? In the second half of his book, Thornton "moves his guns to the other side of the ship," as Hicks (1967) put it (see also Skaggs 2005). We begin to see why Thornton favored a return to the gold standard. Here he emphasizes the need for restraint on the part of the Bank of England, so that its issue of banknotes "vibrates" between an upper and a lower bound. These bounds should cautiously increase over time (PC 259). Limitation of notes requires "some effectual principle of restriction" on discounts when buying bills of exchange. Infamously, the Bank imposed a daily limit on the amount of credit it would extend to the private sector. Thornton is careful not to sound an overly alarmist note, but as the Bank rations credit more tightly, the "pressure" and difficulty of maintaining "regularity of payments" in the capital's money market builds dangerously. Thornton was worried that back on the gold standard, the British government might not see that its long-run aims (such as a stable money supply, gold convertibility, or international payments stability) were best served by periodic departures from the gold standard, followed by a gradual and opportunistic return to that long-run benchmark.

Later writers have argued that the Bank of England and other central banks of the time often rationed credit. Flandreau and Ugolini (2013) note that in 1825 "rampant credit rationing by the Bank of England made major London banks—which were heavily invested in bills—experience a serious maturity mismatch, which forced them to suspend payments." A confluence of the type Thornton feared, leading to credit rationing, appears to have occurred in 1847 (Dornbusch and Frenkel 1984). According to Bignon, Flandreau, and Ugolini (2012), "there was an evolution in the way central banks dealt with crises, from a policy of universal credit rationing before 1850, to a policy that strongly supported the market by providing unlimited loans, or at least much more generous ones." Once Britain returned to gold, it would still face financial crises. Thornton worried that the government

would prioritize maintenance of the gold standard over maintenance of commercial credit, if a domestic credit crisis coincided with an outflow of gold due to capital flight.

Common Shocks, Disparate Responses

The financial crises and inflations of the 1790s caused Jeremy Bentham and Henry Thornton to advocate very different approaches toward stabilizing the banking system. For present purposes, the clearest contrast between these authors lies with the preeminence of liquidity in Thornton's *Paper Credit* as opposed to the role of bank capital in Bentham's *True Alarm*.[8] Thornton admitted that his own bank had held far too little capital as it entered the crisis of 1793, and that "country bankers should be taught . . . to provide themselves with a larger quantity of that sort of property which is quickly convertible into Bank of England notes . . ." (PC 188). Yet in *Paper Credit,* Thornton's spotlight is almost always on liquidity. Bentham never seriously considered liquidity, only "final solvability."

Thornton had a more subtle and informed understanding of the British financial system than Bentham, in part because his own bank occupied an important place in the middle of Britain's credit pyramid. He must have acquired considerable insight into the workings of country banks from his own country bank clientele, while his brother Samuel (who was a Governor of the Bank of England in 1799 and 1801) provided an intimate view into the apex of the system. Indeed, the skill and tact with which Henry Thornton defends the Bank of England has often led readers to mistake him for the Thornton who was a Governor at the Bank. Nevertheless, Thornton's deepest sympathies clearly lie with his fellow bankers in London, who were "pressured" by liquidity panics, contractionary monetary policy, and capital flight. No matter where the pressure originated, it expressed itself in the "general money market." The Bank of England needed to manage this pressure, and if necessary, be aided by loans of government securities to merchants. No matter what the challenge might be, Thornton's solution lies with these actors at the center and not with the merchants and country banks at the periphery of the system. However, he was willing to concede that the Bank of England cannot relieve "every distress which the rashness of country banks may bring upon them: the bank, by doing this, might encourage their improvidence" (PC 188).

Although there is a superficial similarity between Bentham's concern over "universal bankruptcy" and Thornton's "general failure of commercial credit," the two authors are not referring to the same kind of disaster. Bentham's *True Alarm* does

[8]Of course, modern versions of the arguments about the benefits and costs of raising bank capital requirement continue to the present. For example, Admati and Hellwig (2013) advocate greater bank capital (see also Admati et al. 2011). In response, Dewatripont and Tirole (2012) emphasize the cost of additional capital and the alternatives to it—such as contingent convertible bonds and capital insurance. King (2019) alternatively suggests that banks should continually post collateral with their central banks (at predetermined haircuts) sufficient to cover all runnable deposits.

not consider the notion that liquidity provision at the center could either generate or alleviate a collapse of the nation's credit system. This reflects his underappreciation of the interconnections between banks and the market for government debt, and his failure to see how successfully the Bank of England had managed an unprecedentedly large government deficit and debt. Despite these limitations, there can be no doubt that Bentham's advocacy of capital requirements was far ahead of its time, as was his desire to establish a public regulatory agency that would verify bank balance sheets and income statements annually. It is a pity that Bentham's lawyerly attentiveness to bank capital and bankruptcy ended up taking so much longer to enter the mainstream of public discussion, compared with Thornton's sensitive analysis of liquidity. As the better-informed observer of financial affairs, Thornton had the subtler approach. He appreciated the need for a lender of last resort and for flexibility and creativity in response to a crisis.

The main lesson we draw from this episode in the intellectual history of economics is that policymakers need to maintain a balance between prevention of financial crisis and response to that crisis. Bentham was right to insist that without preventive bank regulation, financial crises were inevitable. However, we cannot legislate away what Kane (1988) called the "regulatory dialectic" between regulation and financial industry behavior. If regulations are too binding and supervision not vigilant enough, then banks may find ways to circumvent those regulations in ways that increase systemic risk. A dogmatic reliance on regulations aimed at preventing a crisis could discourage the provision of emergency liquidity because the very possibility of rescue (or "bailout") during a financial crisis may be seen as making such a crisis more likely through moral hazard incentives. Thus, we must hope both for robust Bentham-style regulatory mechanisms to prevent financial crises but also that such preventive mechanisms do not hobble our ability and willingness to offer a Thornton-style response when such crises occur.

■ *We thank David Laidler, Ivo Maes, Joseph Persky, Tim Opiela, and participants of the 2012 and 2014 HES conferences for helpful criticism. Timothy Taylor provided extensive and welcome editorial interventions. We are grateful to all but retain responsibility for all errors and omissions.*

References

Admati, Anat R., Peter M. DeMarzo, Martin F. Hellwig, and Paul C. Pfleiderer. 2011. "Fallacies, Irrelevant Facts, and Myths in the Discussion of Capital Regulation: Why Bank Equity Is Not Expensive." MPI Collective Goods Preprint. (2013/23).

Admati, Anat, and Martin Hellwig. 2013. *The Bankers' New Clothes: What's Wrong with Banking and What to Do About It*. Princeton and Oxford: Princeton University Press.

Anderson, B.L. 1970. "Money and the Structure of Credit in the Eighteenth Century." *Business History* 12 (2): 85–101.

Andréadès, A. 1909. *History of the Bank of England 1640-1903*. London: P.S. King & Son.

Anson, Michael, David Bholat, Miao Kang, and Ryland Thomas. 2017. "The Bank of England as Lender of Last Resort: New Historical Evidence from Daily Transactional Data." Bank of England Working Paper 691.

Antipa, Pamfili, and Christophe Chamley. 2017. "Monetary and Fiscal Policy in England During the French Wars (1793-1821)." Banque de France Working Paper 627.

Ashton, T.S. 1959. *Economic Fluctuations in England, 1700-1800*. Oxford: Clarendon Press.

Ashton, T.S. 1945. "The Bill of Exchange and Private Banks in Lancashire, 1790–1830." *Economic History Review* 15 (1–2): 25–35.

Bagehot, Walter. 1873. *Lombard Street: A description of the Money Market. 3rd ed.* London: Henry S. King & Co. https://oll.libertyfund.org/titles/128 (accessed June 14, 2020).

Bank of England. 2020. "The Bank of England's Balance Sheet 1696-2019." Bank of England Research Datasets. https://www.bankofengland.co.uk/statistics/research-datasets (accessed June 2, 2020).

Bentham, Jeremy. 1838–43. *The Works of Jeremy Bentham,* 11 vols, edited by John Bowring. Edinburgh: William Tait. 11 vols.

Bentham, Jeremy, and Werner Stark. 1952. *Jeremy Bentham's Economic Writings*. London: George Allan and Unwin for the Royal Society.

Bignon, Vincent, Marc Flandreau, and Stefano Ugolini. 2012. "Bagehot for Beginners: The Making of Lender-of-Last-Resort Operations in the Mid-nineteenth Century." *Economic History Review* 65 (2): 580–608.

Bordo, Michael D. and Eugene N. White. 1991. "A Tale of Two Currencies: British and French Finance During the Napoleonic Wars." *Journal of Economic History* 51 (2): 303–16.

Broadberry, Stephen, Bruce M.S. Campbell, Alexander Klein, Mark Overton, and Bas van Leeuwen. 2015. *British Economic Growth, 1270–1870*. Cambridge, UK: Cambridge University Press

Deleplace, Ghislain. 2004. "Monetary Stability and Heterodoxy: A History of Economic Thought Perspective." In *Money, Credit, and the Role of the State. Essays in Honour of Augusto Graziani*, edited by Richard Arena and Neri Salvadori. Aldershot, UK: Ashgate.

Deleplace, Ghislain, and Nathalie Sigot. 2011. "Ricardo's Critique of Bentham's French Manuscript: Secure Currency versus Secure Banks." *European Journal of the History of Economic Thought* 19 (5): 733–64.

Dewatripont, Mathias, and Jean Tirole. 2012. "Macroeconomic Shocks and Banking Regulation." *Journal of Money, Credit and Banking* 44 (S2): 237–54.

Dornbusch, Rudiger, and Jacob A. Frenkel. 1984. "The Gold Standard and the Bank of England in the Crisis of 1847." In *A Retrospective on the Classical Gold Standard, 1821–1931*, edited by Michael D. Bordo and Anna J. Schwartz, 233–76. Chicago and London: University of Chicago Press.

Duffy, Ian P.H. 1980. "English Bankrupts, 1571-1861." *American Journal of Legal History* 24 (4): 283–305.

Flandreau, Marc and Stefano Ugolini. 2013. "Where It All Began: Lending of Last Resort at The Overend, Gurney Panic of 1866." In *The Origins, History, and Future of the Federal Reserve: A Return to Jekyll Island*, edited by Michael D. Bordo and William Roberds, 113–65. Cambridge, UK and New York: Cambridge University Press.

Gorton, Gary B. 2020. "Private Money Production without Banks." NBER Working Paper 26663.

Guidi, Marco E. L. 2010. "Jeremy Bentham, the French Revolution, and the Political Economy of Representation (1788 to 1789)." European Journal of the History of Economic Thought 17(4): 579–605.

Hicks, John R. 1967. "Thornton's Paper Credit." *Critical Essays in Monetary Theory*, 174–88. Oxford: Clarendon Press.

Hollander, Samuel. 1999. "Jeremy Bentham and Adam Smith on the Usury Laws: A 'Smithian' Reply to Bentham and a New Problem." The European Journal of the History of Economic Thought 6 (4): 523–51.

James, John A. 2012. "Panics, Payments Disruptions and the Bank of England before 1826." *Financial History Review* 19 (3): 289–309.

Kane, Edward J. 1988 "Interaction of Financial and Regulatory Innovation." *American Economic Review* 78 (2): 328–34.

King, Mervyn. 2019. "Panel Remarks by Mervyn King on Ten Years after the Great Financial Crisis: What

Has Changed?" Bank for International Settlements Paper 103. 1–3

Marcuzzo, Maria Cristina, and Annalisa Rosselli. 1994. "Ricardo's Theory of Money Matters." *Revue économique* 45 (5): 1251–68.

Montefiore, Joshua. 1803. *A Commercial Dictionary: Containing the Present State of Mercantile Law, Practice, and Custom: Intended for the Use of the Cabinet, the Counting-House, and the Library*. London: Printed for the author. https://catalog.hathitrust.org/Record/009039058/Home. (accessed June 22, 2020).

Neal, Larry. 1994. "The Finance of Business During the Industrial Revolution." In *The Economic History of Britain since 1700*, vol. 1. 2nd ed. Edited by Roderick Floud and Donald McCloskey, 151–81. Cambridge, UK: Cambridge University Press.

O'Brien, Patrick. 2006. "Mercantilist Institutions for the Pursuit of Power with Profit. The Management of Britain's National Debt, 1756-1815." Economic History Working Paper (95/06).

O'Brien, Patrick. 2011. "The Nature and Historical Evolution of an Exceptional Fiscal State and Its Possible Significance for the Precocious Commercialization and Industrialization of the British Economy from Cromwell to Nelson." *Economic History Review* 64 (2): 408–46.

O'Brien, Patrick Karl, and Leandro Prados de la Escosura. 1998. "The Costs and Benefits for Europeans from Their Empires Overseas." *Revista de Historia Económica/Journal of Iberian and Latin American Economic History* 16 (1): 29–89.

Persky, Joseph. 2007. "Retrospectives: From Usury to Interest." *Journal of Economic Perspectives* 21 (1): 227–36.

Pressnell, Leslie Sedden. 1956. *Country Banking in the Industrial Revolution*. Oxford: Clarendon Press.

Ricardo, David. 1951. *The Works and Correspondence of David Ricardo*, edited by Piero Sraffa and M.H. Dobb. Cambridge, UK: Cambridge University Press for the Royal Society. 11 vols. https://oll.libertyfund.org/titles/ricardo-the-works-and-correspondence-of-david-ricardo-11-vols-sraffa-ed. (accessed June 22, 2020).

Schliephake, Eva, and Roland Kirstein. 2013. "Strategic Effects of Regulatory Capital Requirements in Imperfect Banking Competition." *Journal of Money, Credit and Banking* 45 (4): 675–700.

Skaggs, Neil T. 2005. "Treating Schizophrenia: A Comment on Antoin Murphy's Diagnosis of Henry Thornton's Theoretical Condition." *European Journal of the History of Economic Thought* 12 (2): 321–28.

Swagel, Phillip. 2015. "Legal, Political, and Institutional Constraints on the Financial Crisis Policy Response." *Journal of Economic Perspectives* 29 (2): 107–22.

Thornton, Henry, and Friedrich August von Hayek. 1939. *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain*. London: George Allen and Unwin. [1802]. http://oll.libertyfund.org/titles/2041. (accessed June 22, 2020.)

Ventura, Jaume, and Hans-Joachim Voth. 2015. "Debt into Growth: How Sovereign Debt Accelerated the First Industrial Revolution." NBER Working Paper 21280.

Williamson, Jeffrey G. 1984. "Why Was British Growth So Slow during the Industrial Revolution?" Journal of Economic History September 44 (3): 687–712.

Woodford, Michael. 2006. "Comments on the Symposium on Interest and Prices." *Journal of the History of Economic Thought* 28 (2): 187–98.

Journal of Economic Perspectives—Volume 34, Number 4—Fall 2020—Pages 210–217

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 e-mail at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., St. Paul, MN 55105.

Symposia

The *Annals of the American Academy of Political and Social Science* devoted its January 2020 issue to 14 articles on the theme of "Fatal Police Shootings: Patterns, Policy, and Prevention" (https://journals.sagepub.com/toc/anna/687/1). In his essay "Police Killings as a Problem of Governance," Franklin E. Zimring writes: "Police shoot and kill about a thousand civilians each year, and other types of conflict and custodial force add more than one hundred other lives lost to the annual total death toll. This is a death toll far in excess of any other fully developed nation, and the existing empirical evidence suggests that at least half and perhaps as many as 80 percent of these killings are not necessary to safeguard police or protect other citizens from life-threatening force. . . . One reason why U.S. police kill so many

■ *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.34.4.210.

civilians is that U.S. police themselves are vastly more likely than police in other rich nations to die from violent civilian attacks. In Great Britain or Germany, the number of police deaths from civilian attack most years is either one or zero. In the United States—four or five times larger—the death toll from civilian assaults is fifty times larger. And the reason for the larger danger to police is the proliferation of concealable handguns throughout the social spectrum. When police officers die from assault in Germany or England, the cause is usually a firearm, but firearms ownership is low, and concealed firearms are rare. There are, however, at least 60 million concealable handguns in the United States and the firearm is the cause of an officer's death in 97.5 percent of intentional fatal assaults, an effective monopoly of life-threatening force even though more than 95 percent of all assaults against police and an even higher fraction of those said to cause injury are not gun related."

The Spring 2020 issue of *Future of Children* includes nine papers on the theme "How Cultural Factors Shape Economics Outcomes" (https://futureofchildren.princeton.edu/news/how-cultural-factors-shape-economic-outcomes). For example, Ariel Kalil and Rebecca Ryan write about "Parenting Practices and Socioeconomic Gaps in Childhood Outcomes" (pp. 29–54): "Socioeconomic status is correlated across generations. In the United States, 43 percent of adults who were raised in the poorest fifth of the income distribution now have incomes in the poorest fifth, and 70 percent have incomes in the poorest half. Likewise, among adults raised in the richest fifth of the income distribution, 40 percent have incomes in the richest fifth and 53 percent have incomes in the richest half. Many factors influence this intergenerational correlation, but evidence suggests that parenting practices play a crucial role. These include doing enriching activities with children, getting involved in their schoolwork, providing educational materials, and exhibiting warmth and patience. Parental behavior interpreted in this way probably accounts for around half of the variance in adult economic outcomes, and therefore contributes significantly to a country's intergenerational mobility." Daniel Hungerman investigates "Religious Institutions and Economic Wellbeing" (pp. 9–28): "Religious groups discourage unhealthy behaviors and have played an important role in promoting educational attainment and economic wellbeing. Religious participation can increase a person's tolerance of others, and in some circumstances can be particularly beneficial for human capital investments for women. Religion also appears to insure individuals against negative shocks . . . [R]ecent rigorous research suggests that the beneficial effects of religion are often causal, and some work . . . finds that the large association between beneficial outcomes and religion observed in the data may understate religion's true effect."

Edward L. Glaeser and James M. Poterba have edited a collection of eight essays and comments in *Economic Analysis and Infrastructure Investment* (National Bureau of Economic Research, https://www.nber.org/books/glae-6). Gilles Duranton, Geetika Nagpal, and Matthew A. Turner contribute "Transportation Infrastructure in the US." "On average, most US transportation infrastructure is not crumbling, except (probably) for our subways. Over the past generation, the condition of the interstate highway network improved consistently, its extent increased modestly, and

traffic about doubled. Over about the same time period, the condition of bridges remained about the same, the number of bridges increased slowly, and bridge traffic increased modestly. The stock of public transit motor buses is younger than it was a generation ago and about 30 percent larger, although ridership has been about constant. The mean age of a subway car stayed about the same from 1992 to 2017, but at more than 20 years old, this average car is quite old. Subways carry about twice as many riders as they did a generation ago. Speed of travel by car, bus and subway, all declined between 1995 and 2017, most likely as a consequence of large increases in road traffic and subway ridership. Like public transit, the interstate system is largely organized around the provision of short trips in urban areas. . . . Expenditure on transportation infrastructure is growing, and for the most part, allows maintenance to match or outpace depreciation. Moreover, the available empirical evidence does not allow for much confidence in the claim that capacity expansions will lead to economic growth or reduce congestion."

J. Steven Landefeld, Shaunda Villones, and Alyssa Holdren provide a view from the US Bureau of Economic Analysis on "GDP and Beyond: Priorities and Plans," which also includes an introduction by Ernst Berndt and comments by Angus Deaton, Dale W. Jorgenson, Lisa M. Lynch, Paul Schreyer, Louise Sheiner, and Daniel E. Sichel (*Survey of Current Business*, June 2020, https://apps.bea.gov/scb/2020/06-june/0620-beyond-gdp.htm#gdp-nav). Landefeld, Villones, and Holdren write: "In the 1930s, Simon Kuznets, one of the architects of the U.S. national accounts, pointed to the limitations of emphasizing market aggregates, like GDP and national income, and excluding nonmarket activities that have productive value or that enhance economic and social welfare. This criticism is still applicable today. . . . The Bureau of Economic Analysis (BEA) recently embarked on an initiative—"GDP and Beyond"—to identify ways to use its data resources and statistical knowledge to inform the discussion of well-being. . . . In March 2020, the Bureau released a set of prototype measures of economic well-being and growth and prototype estimates of the distribution of personal income. BEA is continuously exploring ways to improve the core GDP accounts, both as a measure of market production and as an indicator of economic well-being and long-term growth, including researching the prices of high-tech goods and services. In addition, the Bureau is updating and expanding its integrated accounts of wealth, productivity, and industry-level production as well as its satellite accounts for sectors like arts and culture, outdoor recreation, health care, and household production. Looking to the future, BEA will turn its attention to longer term projects that require additional research and resources, including valuing "free" digital services, testing alternative aggregate welfare measures, and estimating human capital.

Potpourri

William Easterly considers "Progress by consent: Adam Smith as development economist" (*Review of Austrian Economics*, published online September 10,

2019, https://link.springer.com/article/10.1007/s11138-019-00478-5). "There is a curious notion in development economics that the field emerged out of nowhere right after World War II. I used to share that view . . . It took me embarrassingly long to acknowledge some obvious ancient history of development thinking, and some other development economists are apparently taking even longer. As long ago as 1776, Adam Smith wrote a book called the Wealth of Nations. It turns out, after many hours of careful reading, that the book is indeed about the Wealth of Nations. Far from ignoring the wider world, Smith cited 164 different historical or contemporary place names or names of ethnic groups. . . . The omissions . . . are rare and reflect information availability. Only Australia and New Zealand are left out altogether. Specific place names in Africa are limited to some places on the coast, but there are very important discussions of the African continent as a whole. The rest of the world is well covered . . . Smith has abundant coverage of future Third World places such as Peru, Mexico, Chile, Egypt, India, Africa, Central Asia, and China. Smith's First World success stories are England, lowland Scotland, British North America, and Holland. The future Second World is also covered in discussions of Russia and Eastern Europe. . . . Smith used his widespread examples to test his preferred hypothesis to explain development."

Gary Smith discusses "Data Mining Fool's Gold" (*Journal of Information Technology*, September 2020, pp. 182–194, https://journals.sagepub.com/doi/full/10.1177/0268396220915600). "It is tempting to believe that patterns are unusual and their discovery meaningful; in large data sets, patterns are inevitable and generally meaningless. . . . Data-mining algorithms—often operating under the label artificial intelligence—are now widely used to discover statistical patterns. However, in large data sets streaks, clusters, correlations, and other patterns are the norm, not the exception. While data mining might discover a useful relationship, the number of possible patterns that can be spotted relative to the number that are genuinely useful has grown exponentially—which means that the chances that a discovered pattern is useful is rapidly approaching zero. This is the paradox of big data: It would seem that having data for a large number of variables will help us find more reliable patterns; however, the more variables we consider, the less likely it is that what we find will be useful."

Gavin Wright delivered the Tawney lecture at the Economic History Society meetings on the subject of "Slavery and Anglo-American capitalism revisited" (*Economic History Review*, May 2020, 73:2, pp. 353–83, https://www.ehs.org.uk/app/journal/article/10.1111/ehr.12962/abstract?issue=10.1111/ehr.v73.2, video at https://www.ehs.org.uk/multimedia/tawney-lecture-2019-slavery-and-anglo-american-capitalism-revisited). "To be sure, US cotton did indeed rise 'on the backs of slaves', and no cliometric counterfactual can gainsay that brute fact of history. But it is doubtful that this brutal system served the long-run interests of textile producers in Lancashire and in New England, as many of them recognized at the time. As argued here, the slave South *underperformed* as a world cotton supplier for three distinct though related reasons: the region agreed in 1807 to close the

slave trade and failed to recruit free labourers, making labour supply inelastic; slave owners neglected transportation infrastructure, leaving large sections of potential cotton land on the margins of commercial agriculture; and because of the fixed-cost character of slavery, even large plantations aimed at self-sufficiency in foodstuffs, limiting the region's overall degree of market specialization. These shortcomings in cotton supply had larger ramifications for the course of US development. The slave South became increasingly isolated from the national mainstream, as manufacturers found their most inviting market opportunities in the expanding farm populations and cities of the free states. By the late antebellum period, the slave states emerged as a principal obstacle to the activist growth agenda supported by leading industrial and financial interests. . . . Despite high returns to slave owners, the region underperformed as a cotton supplier, in comparison to a family-farm alternative. As events unfolded, the slave South was neither central nor essential to the mainstream of US economic development."

Joseph E. Aldy and Richard Zeckhauser offer "Three Prongs for Prudent Climate Policy" (*Southern Economics Journal*, July 2020, pp. 3–29, https://onlinelibrary.wiley.com/doi/10.1002/soej.12433). "We've been told, correctly, that the world is running out of time to curb its emission-profligate ways. The world did little mitigation and ran out of the urgent time it was given. And matters have gotten worse, much worse. Emissions cutting, drastic emissions cutting, is still the recommended primary prong of our defense. Experience suggests, and economics reveals, that the magnitude of needed cutting will be almost impossible to achieve in the time available. Moreover, even if the prescribed level of mitigation is met, it may already be too late. A second prong of defense, adaptation, has received some discussion, but very little actual implementation. Adaptation would consist of such measures as building barriers to the ocean, restoring absorptive marshes, repositioning sensitive equipment from cellars to roofs, and preventing new construction in threatened areas. This analysis considers a third prong, amelioration through SRM [solar radiation management] to complement mitigation and adaptation. . . . It would inject aerosols, most likely sulfur particles delivered by airplane, into the upper atmosphere to reflect back incoming solar energy."

The International Comparison Project at the World Bank has published its report "Purchasing Power Parities and the Size of World Economies: Results from the 2017 International Comparison Program" (May 2020, https://openknowledge.worldbank.org/handle/10986/33623). "In 2017, global output, when measured by purchasing power parities (PPPs), was $119,547 billion, compared with $79,715 billion, when measured by market exchange rates. . . . In 2017 lower-middle-income economies contributed around 16 percent to PPP-based global GDP, while upper-middle-income economies contributed 34 percent. At the same time, high-income economies contributed 49 percent. In terms of market exchange rates, these shares were 8 percent, 28 percent, and 64 percent, respectively." The report notes: "ICP PPPs are designed specifically for international comparisons of GDP. They are not designed for comparisons of monetary flows or trade flows. International comparisons of flows—such as development aid, foreign direct investment,

migrants' remittances, or imports and exports of goods and services—should be made with market exchange rates, not with PPPs."

Interviews

"Economics with a Moral Compass? Welfare Economics: Past, Present, and Future," is an interview with Amartya Sen by Angus Deaton and Tim Besley (*Annual Review of Economics*, 2020, 12, pp. 1–21, video also available, https://www.annualreviews.org/do/10.1146/do.multimedia.2020.08.02.01/abs/). "A super-easy target was the so-called compensation tests, which came, oddly enough, from two of the best economists of our time, Nicholas Kaldor and John Hicks. . . . What did it say? To summarize rapidly, consider a change from which some people gain and others lose. If the gainers have gained so much that they can compensate the losers and still retain some gain, then it's an improvement according to the compensation test. So you ask the question, Do they actually compensate the losers? No, they don't have to do it—the losers stay losers (all we are checking is whether they could have been compensated). I mean, what kind of an improvement is that? The losers can rightly think this to be a con job. . . . Why do we need the compensation test at all then, which appears to be either completely unconvincing, or totally redundant? . . . Note that the Bengal famine might have been a compensation test victory, because quite a lot of people gained a lot in 1943, and they could have compensated the new destitutes. They did not have to do it—and the destitutes mostly died—but was there a social improvement there? How could Kaldor, such a fine economist otherwise, propose this criterion? And how could Hicks . . . support it? The answer probably is that both were trying to do welfare economics without having the real courage to go beyond the Pareto principle—without taking on the real problems of distribution, inequality, and poverty. This could not be done, then or now. Happily, both Kaldor and Hicks wrote many other things from which we learn a lot. And each let go of the compensation test, later on."

"Melissa Dell on the Significance of Persistence" in an interview with Tyler Cowen (Medium.com, July 15, audio and transcript, https://medium.com/conversations-with-tyler/melissa-dell-tyler-cowen-history-economic-research-399991533379). "I was presenting some work that I'd done on Mexico to a group of historians. And I think that historians have a very different approach than economists. They tend to focus in on a very narrow context. They might look at a specific village, and they want to explain a hundred percent of what was going on in that village in that time period. Whereas in this paper, I was looking at the impacts of the Mexican Revolution, which is a historical conflict in economic development. And this historian, who had studied it extensively and knows a ton, was saying, 'Well, I kind of see what you're saying, and that holds in this case, but what about this exception? And what about that exception?' And my response was to say my partial R-squared, which is the percent of the variation that this regression explains, is 0.1, which means it's explaining 10 percent of the variation in the data. And I think, you know,

that's pretty good because the world's a complex place, so something that explains 10 percent of the variation is potentially a pretty big deal. But that means there's still 90 percent of the variation that's explained by other things. . . . I'll say the same thing when I teach an undergrad class about economic growth in history. We talk about the various explanations you can have: geography, different types of institutions, cultural factors. Well, there's places in sub-Saharan Africa that are 40 times poorer than the US. When you have that kind of income differential, there's just a massive amount of variation to explain. . . . So there's plenty of room for everybody's preferred theory of economic development to be important just because the differences are so huge."

Discussion Starters

Peter Jaworski has written "Bloody Well Pay Them: The Case for Voluntary Remunerated Plasma Collections" (2020, published by the Adam Smith Institute and the Niskanen Center, https://www.niskanencenter.org/wp-content/uploads/2020/06/BloodyWellPayThem-PeterJaworski.pdf). "The United States is responsible for 70 percent of the global supply of plasma. Along with the other countries that permit a form of payment for plasma donations (including Germany, Austria, Hungary, and Czechia), they together account for nearly 90 percent of the total supply. . . . The United States currently supplies approximately 70 percent of the global need, including about two-fifths of Europe's needs, nearly all of the UK's, over four-fifths of Canada's, over half of Australia's, and around 12 percent of New Zealand's needs for plasma therapies. To put this into perspective, at present, 5 percent of the world's population is responsible for more than half of all the plasma collected in the world. . . . It is no longer reasonable, given the evidence, to continue to insist that non-remunerated plasma collections are able to meet domestic needs, never mind the global needs. Non-remunerated plasma collections have failed everywhere to secure a sufficient supply."

The Congressional Budget Office discusses "Trends in the Internal Revenue Service's Funding and Enforcement" (July 2020, https://www.cbo.gov/system/files/2020-07/56422-CBO-IRS-enforcement.pdf). "The IRS's appropriations have fallen by 20 percent in inflation-adjusted dollars since 2010, resulting in the elimination of 22 percent of its staff. The amount of funding and staff allocated to enforcement activities has declined by about 30 percent since 2010. . . . Between 2010 and 2018, the share of individual income tax returns it examined fell by 46 percent, and the share of corporate income tax returns it examined fell by 37 percent. The disruptions stemming from the 2020 coronavirus pandemic will further reduce the ability of the IRS to enforce tax laws. . . . CBO estimates that increasing the IRS's funding for examinations and collections by $20 billion over 10 years would increase revenues by $61 billion and that increasing such funding by $40 billion over 10 years would increase revenues by $103 billion."

Theresa Levitt discusses "When Lighthouses became Public Goods: The Role of Technological Change" (*Technology and Culture*, January 2020, 61:1, pp. 144–72, https://muse.jhu.edu/article/752963). "The crucial technological change was in the illumination apparatus, with the introduction of mirrors in the 1780s and Fresnel lenses in the 1820s. This was not only a change in technical performance, as each development increased the brightness by more than an order of magnitude. It also brought about the sort of social and institutional transformations that historians of technology have identified as a technological system. As lighthouses became reliably visible at safe distances for sea-coast lighting the first time, their purpose and function changed, as well as their costs and financing. The lighthouse system of the seventeenth century discussed by [Ronald] Coase was fundamentally different from that of John Stuart Mill and Paul Samuelson, with different expectations, expenses, and implications for excludability. While a market could support the lights that existed before 1780, which were primarily effective at close range, it could not support the transformed system that emerged in the wake of improved illumination. Nor could the market provide for the technological improvements, with no private owners of lighthouses investing in Fresnel lenses, one of the key improvements. Only after England introduced greater state intervention did the lights improve."

STUDY WITH **PURPOSE**

*"The one-year, STEM-designated **Master of Arts in International Economics and Finance** prepares professionals to understand advanced economic theories, master professional quantitative and econometric skills, and analyze a wide range of international economics and financial questions."*

— GORDON BODNAR, PhD
Director, International Economics Program and Morris W. Offit Professor of International Finance

JOHNS HOPKINS
SCHOOL *of* ADVANCED INTERNATIONAL STUDIES

LEARN HOW JOHNS HOPKINS SAIS GRADUATES ARE ADVANCING THEIR CAREERS

sais.jhu.edu/mief

WASHINGTON, DC

AEA SUMMER TRAINING & SCHOLARSHIP PROGRAMS

May 27–July 25, 2021 • Washington, D.C.

INTENSIVE TWO-MONTH RESIDENTIAL PROGRAM

TWO LEVELS OF STUDY

COURSE WORK IN MATH, MICROECONOMICS, ECONOMETRICS, AND RESEARCH METHODS

REAL-WORLD EXPERIENCES UNIQUE TO THE NATION'S CAPITAL

PRESENTED IN COLLABORATION WITH:
Women's Institute for Science, Equity and Race, and the Federal Reserve Board as Economics Faculty Collaborators

"My experience with the AEA Summer Program's rigorous curriculum and research project with the guidance of outstanding professors broadened my understanding of what it means to be a researcher and connected me with a powerful network of aspiring economists and economists of color."

Fanta Traore
HU (B.A. '15)
AEASP 2017

"It is inspiring to see the caliber of research the students are able to produce by the end of the program."

Jevay Grooms
AEASP 2016 and 2017
AEASP Research and Faculty Fellow

APPLICATION DEADLINE: January 31, 2021

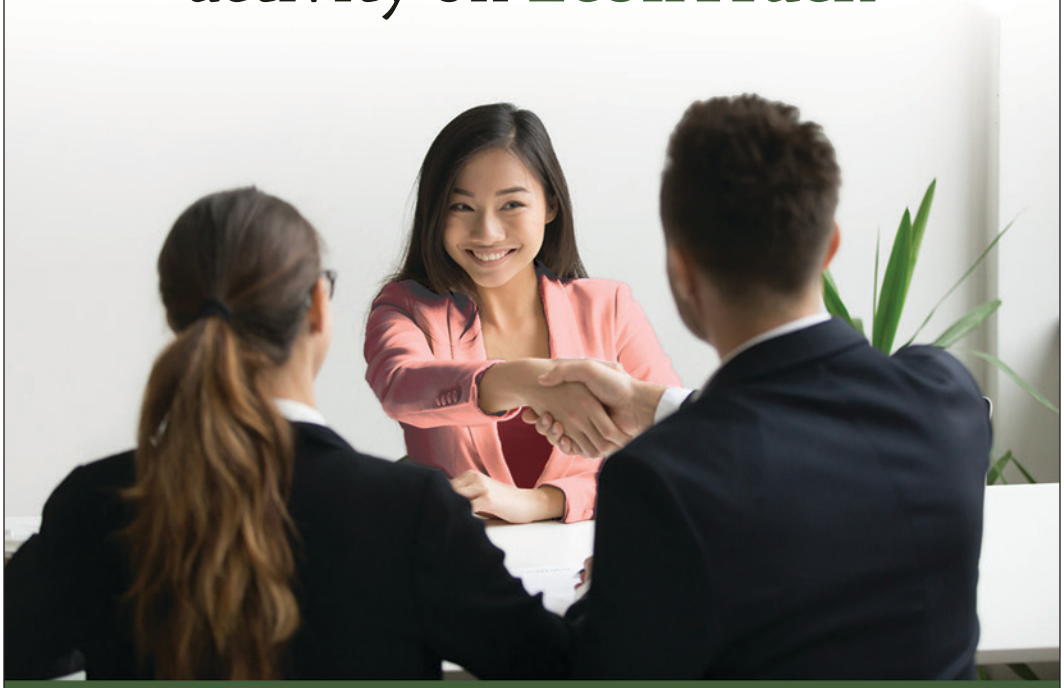
FOR MORE INFORMATION:
E-mail aeasp@howard.edu
http://economics.howard.edu/aeasp

Women's Institute for Science, Equity and Race

Association for Economic Research of Indigenous Peoples

▶▶▶▶ *Did you know?*

You can monitor job market activity on **EconTrack**



All information is provided directly by employers.

www.aeaweb.org/**econtrack**

- Job application deadlines
- Scheduled interviews
- Scheduled campus visits
- List of campus invitees
- When offers are extended and accepted

The *Journal of Economic Perspectives*: Proposal Guidelines

Considerations for Those Proposing Topics and Papers for *JEP*

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.

Philosophy and Style

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of *JEP* articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author.** The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

JEP does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-U.S. country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives.* Our stock in trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship. If you are not familiar with this journal, it is freely available on-line at www.aeaweb.org/journals/jep.

Guidelines for Preparing *JEP* Proposals

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP.* However, given

the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion.
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and authors an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future articles for *JEP* may be sent (electronically please) to the assistant managing editor, Alexandra Szczupak at a.szczupak@aeapubs.org. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

Guidelines for Empirical Papers Submitted to *JEP*

JEP is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1) The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.

2) In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.

3) The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.

New AEA Training Opportunity!

EDUCATE Workshop

EXPANDING DIVERSITY in UNDERGRADUATE CLASSES with ADVANCEMENTS in (the) TEACHING (of) ECONOMICS

Overview

This workshop provides opportunities for instructors of introductory courses to take part in course design activities and experience pedagogical strategies that will engage all of the students they teach. Attendees will have opportunities to identify learning objectives that focus on the students' ability to "do economics" and to participate in pedagogical practices that enable students to be active participants in economic analysis.

All accepted applicants are expected to fully engage with each of the three phases of the overall program including constructing learning objectives, studying pedagogical practices that are collaborative and inclusive including cooperative learning, engaging lectures, data integration, and classroom experiments, and integrating lessons learned into their own courses. Participants will be provided opportunities to share their work at the 2022 CTREE and ASSA meetings.

Participants will engage with issues of diversity and inclusion throughout the workshop including opportunities to think critically about course goals and learning outcomes, their relationship to pedagogical choices and assessment, and how such decisions might have disparate effects on those of different races, genders, and ethnicities.

Eligibility

Applicants must be scheduled to teach introductory microeconomics, macroeconomics, or a combined course in the spring of 2021 and also plan on teaching one of these courses in the fall of 2021. Preference will be given to those less than 6 years since PhD. To meet the goals of increasing diversity in the profession, the first EDUCATE cohort will be chosen to represent a diverse set of institutions and instructors.

Cost

Accepted applicants must make an electronic payment of $100 to confirm their spot in the workshop. Those who complete and submit a final post-workshop implementation plan (summer 2021) will be provided a $500 stipend.

The application portal is open now, with a rolling acceptance procedure starting on October 1, 2020 and continuing until all workshop slots are filled. Workshop details and the application portal are available at https://www.aeaweb.org/go/educate-workshop.

EDUCATE
Virtual Workshop
(A synchronous Zoom and Canvas supported course)
January 5-6-7, 2021

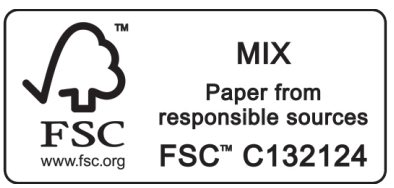
For more information go to
https://www.aeaweb.org/go/educate-workshop

EDUCATE is sponsored by the AEA Outreach Task Force and the AEA Committee on Economic Education and is presented in conjunction with the AEA 2021 Continuing Education Program.

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 access to JEP articles, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $24.00, $34.00, or $44.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

The following Statement of Ownership, Management and Circulation is provided in accordance with the requirements, as contained in 39 U.S.C. 3658. Journal of Economic Perspectives is owned, managed and published by the American Economic Association, a nonprofit educational organization, located at 2014 Broadway, Suite 305, Nashville, Davidson County, TN 37203. The Editor is Enrico Moretti, UC Berkeley. The Managing Editor is Timothy Taylor: Journal of Economic Perspectives, 2403 Sidney Street, Suite 260, Pittsburgh, PA 15203. The tax status of the American Economic Association has not changed during the preceding twelve months. During the preceding twelve months, the average number of copies printed for each issue was 4,408; the average total paid and/or requested circulation, 3,766; the average total non-requested distribution, 0; the average number of copies not distributed, 642; the average total distribution, 3,766. Corresponding figures for May 2020, the issue published nearest to filing date: total number of copies printed, 4,119; total paid and/or requested circulation, 3,569; total non-requested distribution, 0; number of copies not distributed, 550; total distribution, 3,569. During the preceding twelve months, the average number of requested and paid electronic copies of each issue was 939; the total average requested and paid print and electronic copies, 4,705. Corresponding figures for May 2020, the issue published nearest to filing date: number of requested and paid electronic copies, 927; the total requested and paid print and electronic copies, 4,496. Certified by Barbara Fiser, Director of Finance and Administration.



The Journal of

Economic Perspectives

Fall 2020, Volume 34, Number 4

Symposia

How Much Income and Wealth Inequality?

Economics and Epidemiology

Articles

Features

AMERICAN ECONOMIC ASSOCIATION