The Journal of Economic Perspectives

A journal of the American Economic Association

Winter 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), Winter 2020, Vol. 34, No. 1. The JEP is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $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 1 • Winter 2020

Symposia

Articles

Feature

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 3–30

Dynamism with Incommensurate Development: The Distinctive Indian Model

Rohit Lamba and Arvind Subramanian

Constituting one-seventh of humanity, fissured horizontally by region, religion and language, and ossified vertically by caste and patriarchy, India is as much a subcontinent of quasi-sovereign states as a unitary country. Against this background, the paper explores some of the puzzles and anomalies that have characterized India's development, with a focus on the period since 1980. Its theme is the contrast between India's growth dynamism—notably rapid, long, and consistent—and its social and structural transformations, which although tangible and substantial, have not matched its overall growth.

When India gained independence in 1947, it was a poor country with per capita GDP of $820 (in constant 2011 US dollars at the purchasing-power parity exchange rate). More than 70 years later, India's per capita GDP is approximately $6,500, making it a lower middle-income country. Concomitantly, the poverty rate has declined from about 70 percent to 21 percent (in the most recent official statistics in 2011), and the child mortality rate has fallen from 30 percent to 5 percent. Meanwhile, life expectancy has increased from 32 years to nearly 70 years and the primary school completion rate from 40 percent in 1971 to nearly 100 percent today (based on data from the World Bank, Ministry of Finance 2017, and the Maddison Project). In this essay, we begin with a brief overview of India's economic growth since independence. In particular, three phases of steady market-friendly reforms in the 1980s,

■ *Rohit Lamba is Assistant Professor of Economics, Pennsylvania State University, University Park, Pennsylvania. Arvind Subramanian is Visiting Lecturer in Public Policy, Harvard Kennedy School, Cambridge, Massachusetts. From 2014 to 2018, he served as Chief Economic Adviser to the Government of India. Their email addresses are rlamba@psu.edu and arvind_subramanian@hks.harvard.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.3.

1990s, and 2000s have helped create a rare growth dynamism. We show that in the post-World War II period, India belongs to a small group of countries that have grown considerably and sustainably for a long period of time. But India's macroeconomic road has not been without bumps, including a conventional balance-of-payments crisis in 1991 and an exposure of macroeconomic vulnerabilities in the wake of the "taper tantrum" in 2013. The 2010s have been marked by the "twin balance sheet" crisis—an overindebted corporate sector and a public sector-dominated baking system laden with non-performing assets—which has slowly if undramatically corroded the dynamism in growth.

We then turn to India's unusual sequencing of economic and political development. In India, democratization preceded development, which is a pattern unlike the successful economic models of the East Asian countries or of the West. With rapid growth over the last four decades and a long-standing democratic political system, one would expect that key indicators of India's development ought to improve appreciably. For example, India would be expected to structurally transform its economy from one reliant on agriculture to one reliant on high-wage, high-productivity manufacturing; to reduce regional disparities as the fruits of growth spread from richer to poorer geographies; to narrow differences across social and religious groups; to reduce discrimination against women and facilitate the entry of women into the workforce; to solve chronic malnutrition amongst children; and to improve the environment. But as we will discuss, such improvements have either not occurred or at least not in proportion to India's apparent dynamism.

One possible explanation of why India's overall economic dynamism has been accompanied by incommensurate development is that India's GDP growth has been overestimated. There are a number of signs that when India revamped its methods of measuring GDP in 2011, it may have done so in ways that led GDP growth to be overestimated by about 2.5 percentage points per year. However, given that India's annual estimate growth rates of per capita GDP have exceeded 6 percent since 2002, even subtracting a couple of percentage points would still mean that India's growth was quite rapid over the last four decades. In the conclusion, we offer some reflections on the future of this distinctive Indian model of economic growth and its development consequences.

Growth Dynamism

For the first few decades after India gained independence from British rule in 1947, its public sector seized the "commanding heights" of the economy, while the private sector was allowed to operate in "nonessential" sectors.[1] Import-substituting industrialization was pursued through sweeping controls on imports, foreign direct investment, and foreign technology. In a distinctively Indian twist, and offsetting these

[1] A timeline of India's major economic policy choices is presented in the online Appendix available with this paper at the *Journal of Economic Perspectives* website.

protectionist policies which targeted foreign supplies and firms, homegrown entrepreneurship was "taxed" through extensive controls on domestic private production and capacity, and in 1969, domestic private sector banks were nationalized (Bhagwati and Desai 1970; Joshi and Little 1996). The British raj had been exchanged for a "license-quota-permit raj." This Kafkaesque maze of controls contributed to India's unexceptional "Hindu rate of growth" (as it was often called) of 1.4 percent per capita between 1950 and 1980.

In the 1980s, in its first phase of economic reforms, India started moving away from this model by implementing modest pro-business reforms (Kohli 2010). The changes favored domestic producers and incumbents by relaxing constraints on them and easing their access to capital, inputs, and technology, but without exposing them to greater competitive threat. To use a phrase from Qian (2017) in the context of China, it was a model of "reforms without losers." The early modest reforms elicited a large productivity response, perhaps in part because they signaled an attitudinal shift from the government, and in part because India was so far from its income-possibility frontier (Rodrik and Subramanian 2005). India's GDP growth more than doubled in the 1980s to a new trajectory of about 3.5 percent per capita a year. However, macroeconomic profligacy ensued as India's public expenditure and fiscal deficit rose substantially, culminating in a major balance-of-payments crisis in 1991. The crisis was "Hirschmanian" in that it was deep enough to legitimize sweeping and politically costly reforms but not so deep as to wipe out the fiscal or political means to make them (see Adelman 2013 for a description of this idea by Albert Hirschman).

Thus, in the second and perhaps the most decisive phase of reforms in the early 1990s, India responded by repudiating the dirigiste past: it introduced pro-market, pro-competition policies, liberalizing foreign trade, exchange rate, capital and investment controls, as well as domestic private-production regimes (for discussion and details, useful starting points are Bhagwati and Srinivasan 1995; in this journal, Ahluwalia 2002; DeLong 2003; Bhandari and Lamba 2016; Mohan 2018; Sitapati 2018).

A third phase of reforms followed in the early 2000s as more sectors were opened to competition, and financial liberalization and other tax and regulatory reforms were undertaken (Panagariya 2008). These changes, combined with favorable external conditions, propelled India into a boom phase, wherein per capita growth has averaged 6.2 percent since 2002.

The mid-2000s also witnessed a surge in redistribution through rights and entitlements to food, rural employment, and education. This reflected both increased fiscal ability as revenues surged with growth and a desire to spread the benefits of growth, especially when the capacity for provision of public goods such as health remained weak (Dréze and Sen 2013). As a result of economic growth and these policy changes, millions of Indians have been pulled out of poverty, and a sizable middle class has emerged.[2]

[2]See Roy (2011) for an overview of Indian economic history under British rule and Basu (2018) for a summary post-Independence. See also Bardhan (1999) on the political and social constraints on development in India in the twentieth century.

Table 1
Countries Matching or Surpassing India's Pace, Duration, and Stability of Growth

Country	*Average growth rate (%)*	*Duration (years)*	*Maximum 38-year growth rate (%)*	*Takeoff year*	*Average Polity score*
Botswana	6.4	59	6.6	1959	7.0
Singapore	6.3	60	6.4	1958	2.6
Republic of Korea	6.2	60	6.9	1957	4.2
Taiwan	6.2	66	6.7	1951	3.6
Malta	5.5	60	6.0	1958	N/A
Hong Kong (SAR)	5.4	60	5.4	1952	N/A
China	5.4	49	6.2	1969	0
Thailand	4.8	43	5.4	1955	2.8
India	4.6	38	4.6	1980	8.6
Malaysia	4.5	39	4.5	1959	6.1

Source: Maddison Project Database (2018) and the Polity IV dataset.

No discussion of India is complete without a comparison to China, its equally large and complex neighbor. While both countries have done remarkably well in pulling hundreds of millions out of poverty over the last four decades, China has done so at a brisker pace and attained a much higher level of per capita GDP, thereby spawning a bigger middle class. Even so, the dynamism of the Indian growth story puts it in a small group of post-World War II economies, which have sustained a comparable level and pace for a significant period of time.[3]

Since its growth takeoff in 1980, India's growth of GDP per capita has averaged 4.6 percent for 38 years from 1980–2018, with no decadal average during this 38-year period falling below nearly 3 percent. In Table 1, we report all countries since 1950 that (1) have grown at 4.5 percent or more for at least 38 years in this period and (2) during which any consecutive 10-year average has not fallen below 2.9 percent. Only nine countries make the cut. Seven of those are in East Asia and one each in sub-Saharan Africa and Europe. Among these countries, India is the outlier in terms of political freedom, with only Botswana coming close to being a persistent democracy in this period of high growth.

Of course, one can tweak this comparison in a number of ways. For example, if we relax the second criterion of ten-year averages, we notice that Japan's growth turns out to be volatile in the 1970s (2.26 percent for ten years starting in 1974) and Vietnam just misses the growth criterion because it grew at –3.5 percent in the year 1979–1980 (2.29 percent for ten years starting in 1980). One should also

[3] In this journal, Bosworth and Collins (2008) provide a comparative analysis of the China-India growth story.

note that countries that start poorer have a greater ability to grow faster. But whatever precise metric is chosen, India's growth performance since 1980 has been unusual in pace, duration, and nonvolatility, facilitating a fourfold increase in average living standards.

As India's economy moved to a faster rate of growth, the state could not step up its regulatory role. Major corruption scandals erupted in the allocation of natural resources such as spectrum, coal, and land, a "rents-raj" emerged as the twenty-first century analogue of the earlier "licence-quota-permit raj" (Rajan 2012). India's infrastructure boom of the 2000s came to be associated with dubious lending from public sector banks to private corporate houses (Crabtree 2018). The accumulated experience of corrosive links between the state and private capital has led to "stigmatized capitalism," which undermines the legitimacy of both actors (Subramanian 2018).

This overexuberant and tainted financing has also bequeathed a toxic legacy of fragile, overindebted corporate sector balance sheets and counterpart nonperforming assets in the financial system, especially the public sector banks—the "twin balance sheet" problem (Ministry of Finance 2015, 2017 in chapter 1 and chapter 4, respectively; Rajan 2018). A new bankruptcy code has been adopted in an attempt to facilitate the resolution of bad assets, but it is still too early to evaluate its effectiveness.

Since 2014, India's government has embarked on a "new basic needs welfarism." Its affirmative agenda involves the state providing essential private goods and services to the poor such as bank accounts, cooking gas, housing, toilets, power, and emergency medical insurance. This welfarism leverages financial inclusion, biometrics, and mobile technology (referred to as the "JAM" trinity in India) to build state capacity, which in turn can more effectively deliver the benefits (George and Subramanian 2015). It is unusual in its scope because it still excludes effective provision of public goods such as health and education—a longstanding failing of Indian polity and society.

A major fiscal and efficiency-enhancing reform was the implementation in 2017 of a national Goods and Services Tax. Its likely benefits are threefold: eliminating the multiplicity of taxes across the Indian states and creating a simple, common indirect tax system; reducing the transaction costs of trading across states and transforming India into a common market; and exploiting the self-policing nature of the valued-added tax to reduce evasion, improve compliance, and strengthen governance (Adhia and Subramanian 2016).

On the other side, a controversial demonetization policy in November 2016 withdrew 86 percent of the currency in circulation, a monetary shock that imposed large costs especially on the informal sector reflected in the increased take-up of the employment-guarantee scheme (Ministry of Finance 2017, chapter 1). Puzzlingly, the impact of demonetization on the formal economy has been less adverse than anticipated (for discussions on demonetization, see Chodorow et al. 2018 and the article by Lahiri in this symposium).

Politics and Economics: India's Unusual Sequence

Political institutions and economic development are strongly correlated. One direction of causation owes to the "modernization hypothesis" of political science: the empirical regularity that countries start democratizing as their incomes grow and sustain democracy only at higher levels of income.[4] India has famously defied this hypothesis. Political scientists often describe as an anomaly how India has managed to sustain a democracy under inhospitable conditions of low income and literacy, a predominant rural economy, and major social cleavages—especially once the factor of caste is taken into account. Varshney (1998) provides a thoughtful analysis of democracy in India as a puzzle for most standard theories of political economy. Figure 1 plots a score of democracy on the *y*-axis and GDP per capita at independence and ethnic fractionalization on the right and left panels of the *x*-axis, respectively. It is striking how few uninterrupted democracies there are in the post-World War II period, and India stands out for having sustained democracy despite being poorer, more fractionalized, and of course much larger.

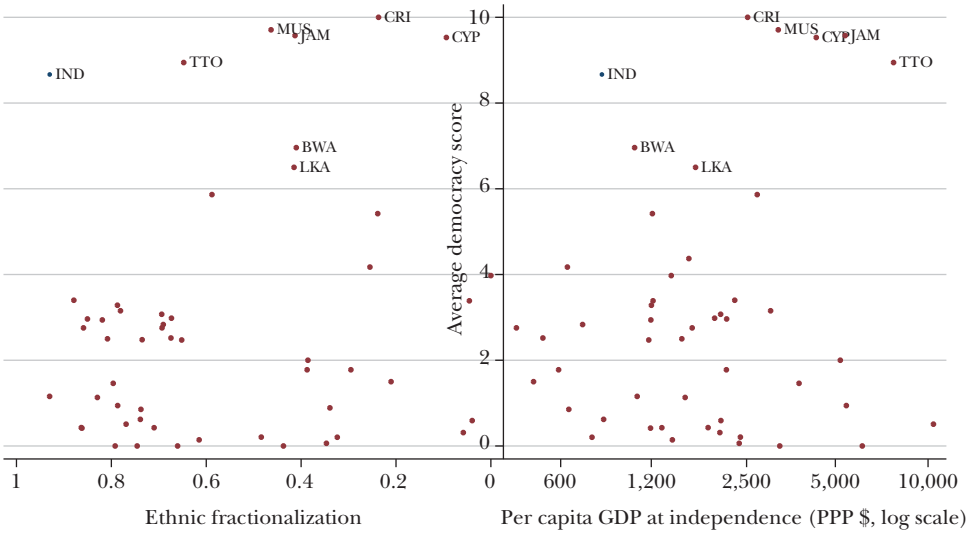
The other direction of causation is associated with the work of North (1990), Acemoglu and Robinson (2012), and Acemoglu et al. (2019), among others. This suggests that democratic political institutions causally affect growth by creating a conducive climate for investment and absorption of new technology and by helping to prevent the stifling entrenchment of vested interests and monopolies.

Before we delve deeper into these links for India, we first establish the common empirical patterns of the sequencing of economic and political development. Historically, if one looks at successful economic transitions, there are really two models in terms of the pace and sequencing of economic and political development. In the first model, comprising Europe and North America and starting with the Industrial Revolution, the economic transition occurs gradually over time with political development, especially suffrage, evolving alongside (Engerman and Sokoloff 2005). The combination, in other words, is one of steady economic growth (about 1.5 percent for nearly 200 years) along with steady political development.

This model of development is exemplified by the United States and the United Kingdom in Figure 2. It plots a democratic index against per capita GDP in the time period 1810–2015. The number in square brackets indicates the average growth rate over that entire time. The big dots for the United States and United Kingdom show the "development time" (that is, the path taken by a variable as the underlying per capita GDP changes) at which the country completed its path to universal suffrage, defined here for practical purposes as the date that the right to vote was granted to women. By this definition, the United States provided its citizens with universal

[4]Lipset (1959) first characterized the modernization hypothesis, and Huntington (1969) and Fukuyama (1989) built on these ideas: the latter proclaiming its much-cited apogee, the so-called "end of history." There is now some evidence that the regularity has been weakening over time, such that "the link between economic development and what is generally called liberal democracy is actually quite weak and may even be getting weaker" (de Mesquita and Downs 2005). Acemoglu et al. (2009) question the empirical validity of the modernization hypothesis.

Figure 1
Democracy and Initial Conditions: Income and Fractionalization



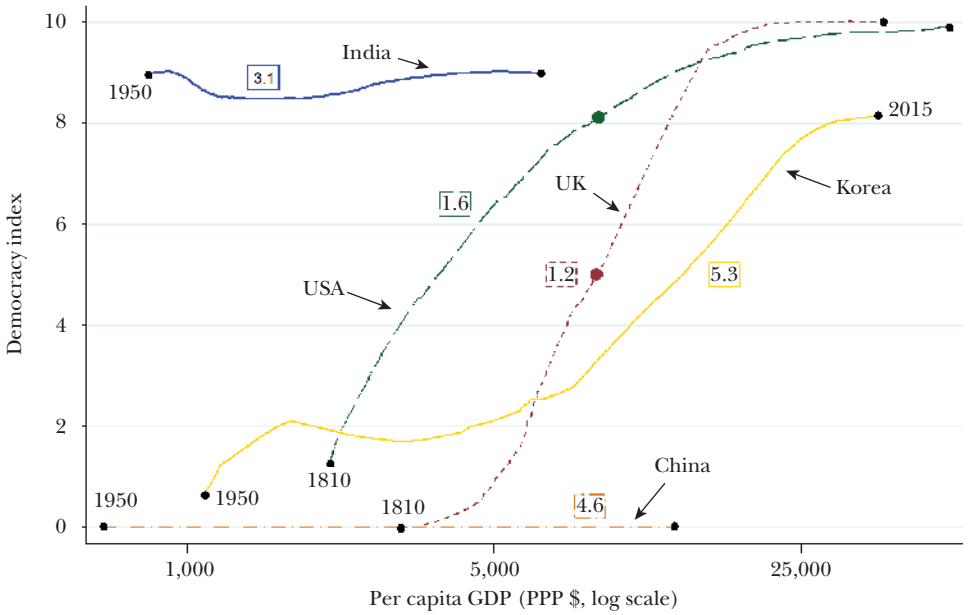
Source: The measure for ethnic fractionalization is taken from Alesina et al. (2003). The number for India is updated using Banerjee and Somanathan (2007). The year of independence for countries comes from ICOW Colonial History data, version 1.1. The average democracy score is calculated using Polity IV from the year of independence to 2015. Costa Rica gained independence in 1821, but the income data is available from 1920, so we use numbers from 1920.
Note: The figure plots an average democracy score since the time of independence (*y*-axis) against a measure of ethnic fractionalization on the left panel and per capita GDP on the right panel (*x*-axis).

suffrage in 1920, the United Kingdom in 1928, France in 1944, and Switzerland in 1971.

The other model of successful transition to upper middle-income and high-income status is represented mostly by East Asian countries, which started off as authoritarian political regimes controlled by the military (South Korea), party (China and Taiwan), monarchy (Thailand), or an individual (Indonesia). In such cases, economic growth has been more rapid during the post-World War II period (an average annual rate of more than 4.5 percent for both China and Korea), while political openness has either sluggishly followed economic development (as in South Korea) or still remains limited (as in China). Again, Figure 2 plots the comovement of democracy and growth for South Korea, China, and India from 1950–2015: South Korea reached the level of economic development of the United Kingdom in half the calendar time, and China is catching up fast in economic terms without much expansion in political freedom.

India's story has been different from both these models in both respects. First, India's pace of economic growth since World War II (an average annual rate of

Figure 2
Patterns of Sequencing of Economic and Political Development



Source: For India, China, and Korea, the democracy score is directly reported from Polity IV; the same is true for the United States and United Kingdom after universal suffrage. Before suffrage for the United States and United Kingdom, the weighted democracy score in year t is Democracy score × Voter participation in year t/Voter participation just after universal suffrage. For this latter construction, voter participation is recorded from the Polyarchy dataset. For computing the weighted democracy score before suffrage, ideally, we would want the fraction of population who has voting rights, but since that information is not available, the voter-participation rate is used as a proxy. Finally, LOWESS (Locally Weighted Scatterplot Smoothing) is used to generate the curve of moving averages.
Note: The figure plots a democracy score (y-axis) against per capita GDP (x-axis).

3.1 percent) has been more rapid than the steady pace of North America and Western Europe, but less so than the dynamic East Asian economies. More strikingly, India's political development has not proceeded alongside or after economic growth, but instead, preceded the economic transition, reflected in the grant of universal franchise in one stroke immediately after independence (Guha 2007). In Figure 2, India stands out for starting with a high democratic score that was only achieved much later in "development time" by the United States and United Kingdom and that remains elusive for the East Asian economic successes, especially China.

In short, combining Figures 1 and 2, we see that India has defied the modernization hypothesis with democratization occurring before development and despite deep ethnic cleavages, while China defies the modernization hypothesis in the

other direction by rapidly striding towards development with little progress on democratization.

Incommensurate Development

Rapid overall growth in GDP and a sustained democracy should be accompanied by development across a number of dimensions. However, India's broader pattern of development has not matched its overall economic growth along a number of dimensions discussed in this section: sectoral composition of growth and employment, distribution across geography and by caste and religion, progress on gender equality and children's nutrition, and mitigation of rising environmental risks.

Premature Deindustrialization and Precocious "Servicification"

India's growth dynamism has been associated with an unusual structural transformation, as first discussed in Kochhar et al. (2006). Herrendorf, Rogerson, and Valentinyi (2014) provide a detailed theoretical and empirical overview of the current thinking on structural transformations. In their spirit, we present four facts.

First, India has, atypically, skipped the low-skilled manufacturing stage and proceeded straight to services; we call this "premature deindustrialization, precocious servicification." Second, this pattern reflects and is perhaps caused by a deeper misallocation of factors of production, based on exploiting comparative advantage in scarce skilled labor rather than abundant unskilled labor. Third, despite unusual specialization and a history of restrictive trade policy, India is a fairly open economy, much more than expected given its level of development and size.[5] Fourth, the misallocation of physical and human capital entails significant distributional costs. As a result, many people in India are not sharing the fruits of its growth.

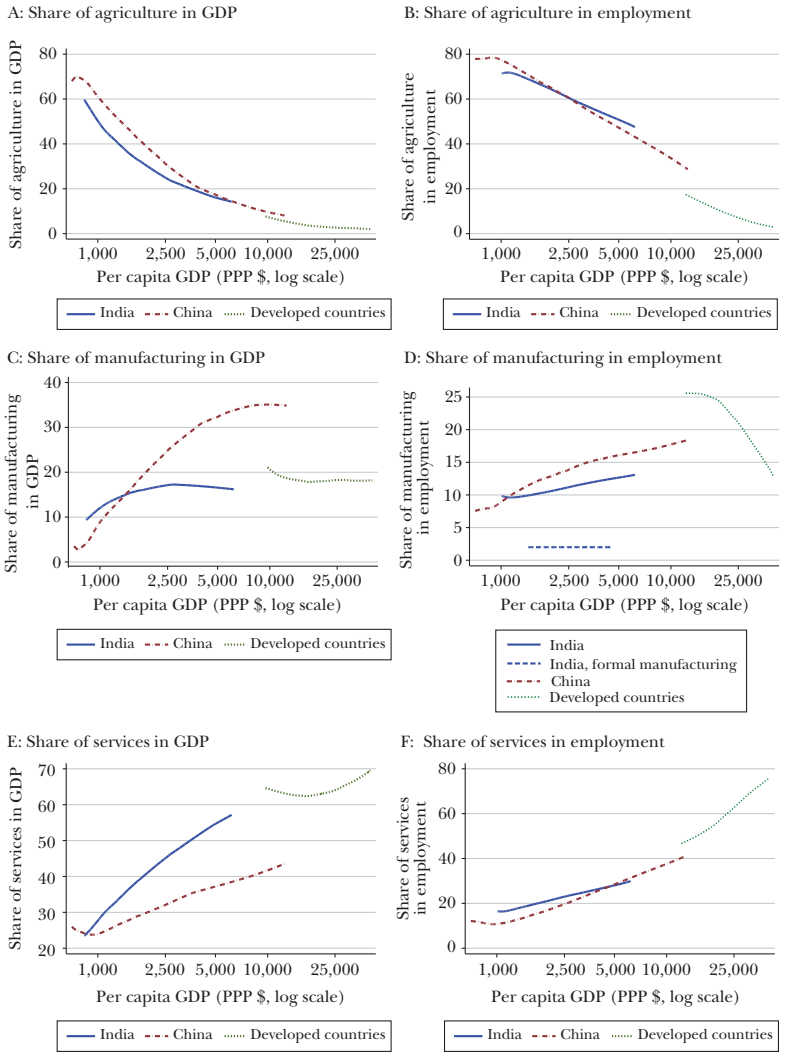
Figure 3 plots in development time the share of agriculture, manufacturing, and services in total output (left-hand panels: A, C, and E) and in total employment (right-hand panels: B, D, and F) for India, China, and developed countries as a group. The period covered is 1950–2017.

The trajectory of agriculture for all three is fairly similar, though for manufacturing and services there is a sharp contrast between India and the others. India's share of manufacturing in GDP is always well below that of China at comparable levels of development. India's share peaks at 19.2 percent at a per capita GDP of $2,177 in the year 1996; in contrast, China's share of manufacturing in GDP peaks at 36.5 percent at a per capita GDP of $9,555 in the year 2010 (at constant US dollar prices). The share of manufacturing of today's advanced countries has always been greater than India's and had also peaked at a higher level in their development process. Moreover, even within manufacturing, the share of formal manufacturing

[5] This "trade puzzle" is discussed in detail in the online Appendix.

Figure 3

Share of Different Sectors in GDP and Employment over Development Time for India, China, and the Developed Countries



Source: The share of the respective sectors is taken from the GGDC 10-Sector Database. Since the GGDC data ends in 2011–2012, it is augmented till 2017 using the WDI database for India and China. To make two datasets comparable, the mean values for the share of agriculture, manufacturing, and services for years 2006–2012 from both datasets are computed and then the WDI numbers after 2012 are updated by dividing them with the WDI mean and multiplying by the GGDC mean. The developed-country average for the share in agriculture, manufacturing, and service is constructed using the GGDC dataset by taking the simple average for the United States, West Germany, Spain, France, the United Kingdom, Italy, the Netherlands, Sweden, and Japan. If the data is not reported for some country in a given year, it is removed from the simple average.

Note: As in Figure 2, we use the LOWESS method to smoothen the curves. The figure plots the share of agriculture, manufacturing, and services in GDP and employment (*y*-axis) against per capita GDP (*y*-axis).

in India is extremely small, as shown in Figure 3, panel D. Overall, India is now a classic case of "premature deindustrialization" (Rodrik 2016).

For India, the flip side of premature deindustrialization is precocious servicification of its economy, as Figure 3, panel E shows. India's services share is consistently greater than China's and is on pace to reach that of advanced countries, but at much lower levels of per capita GDP (Amirapu and Subramanian 2015). Traditional theories have placed a hierarchy on the "natural" order of economic development: first a structural transformation from agriculture to low-skilled manufacturing, then the next transformation to high-skilled manufacturing, and eventually services.[6] India has turned this theory on its head by leapfrogging manufacturing and adopting a low and high skill-intensive services transformation. India has thus grown by defying, rather than deifying, its comparative advantage in abundant unskilled labor.

These domestic patterns of specialization and the revealed comparative advantage have trade counterparts. Premature deindustrialization and precocious servicification reflect weak and strong international competitiveness of the respective sectors. In a comparison with countries that have had a growth rate of at least 4.5 percent over 30 years in the post-World War II era, India's manufacturing exports/GDP ratio peaked at 10.5 percent compared to 32.5 percent for China, 71.2 percent for Vietnam, and 18.1 percent for Bangladesh. In contrast, India is amongst the best performers in this group on the metric of peak skill-intensive exports to GDP ratio (at 7 percent), bested only by Singapore and Hong Kong. In this peer group of fast growers, India has failed to exhibit competitiveness in manufacturing while displaying it in skill-intensive services.

Could India's premature deindustrialization be explained by bad timing? The answer seems to be negative because India does worse on manufacturing than all three vintages of growth stars: Singapore, Hong Kong, and Thailand, which started accelerating in the 1960s and 1970s before India; China, whose growth acceleration was contemporaneous with India's; and Indonesia and Vietnam, whose growth acceleration started about a decade after India. For all these countries, both manufacturing shares in GDP and manufacturing export shares have been greater.

Accounting for India's unusual pattern is a combination of policy and chance, which de facto converted a country physically abundant in unskilled labor into one that was competitively scarce in it. The "license raj" created a web of incentives and disincentives that not only raised the cost of unskilled labor but militated against entry of new firms and employment expansion (Kochhar et al. 2006; Hsieh and Klenow 2009). Formal manufacturing suffered and export opportunities were thwarted. When the information technology revolution came along in advanced countries in the 1990s, India was well situated to exploit the opportunities because of its pool of skilled, English-speaking labor, which in turn was a legacy of the early

[6] Clark (1940) and Kuznets (1957) are some of the early references here. Ray (2010, in this journal) emphasizes the role of structural change more generally through the interaction of sectoral shifts in allocation of labor and capital with technological progress and its consequences for the distribution of income.

Nehruvian emphasis on higher education. It was also crucial that the new service sectors escaped the stifling reach of the license raj, echoing the famous quip by Gurcharan Das (2012) that India grows at night when the government sleeps.

India's unusual specialization-cum-transformation has had distributional consequences captured in Figure 3, panels B, D, and F, which show sectoral employment shares for India, China, and developed economies. At the extensive margin, a reallocation of labor from low- to high-productivity sectors increases growth and improves distribution in the economy (McMillan, Rodrik, and Verduzco-Gallo 2014). At the intensive margin, if initially more productive sectors have a higher growth rate of productivity than the initially low-productive ones and there is no reallocation of labor, distributional costs are exacerbated. These individual effects and their interaction determines whether the growth is equitable or uneven.

The contrast between India and other countries in their divergent structural transformations is striking. India's employment share of manufacturing and services (Figure 3, panels D and F) is much lower at a comparable income level; worse, high-productive formal manufacturing is even smaller (dotted line at bottom of Figure 3, panel D). Thus, India's dynamic and high-productivity activities have benefited a small fraction of the workforce.[7] The counterpart of this is the continuing high share of labor still employed in agriculture characterized by anemic growth in productivity. The Lewis (1954)-style transformation of labor moving out of low-productivity agriculture in large quantities has still not happened in India, with adverse consequences for income distribution.

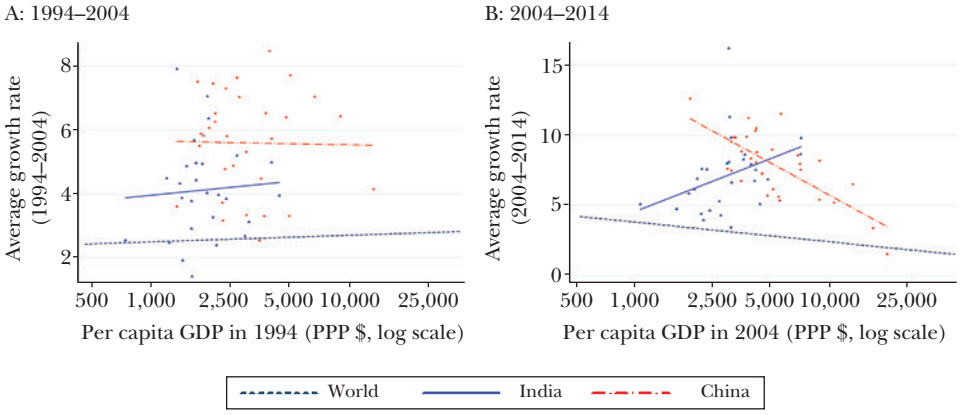
In short, India's path of specialization has been unusual: premature deindustrialization and precocious servicification, combined with weak agricultural productivity and a lack of reallocation of employment away from low- to high-productivity sectors. This path carries the risk that patterns of inequality will persist and may even worsen over time.

Spatial Divergence and Inequality

As a large and diverse country, achieving balanced regional and spatial growth has been a long-standing policy goal of the Indian state. In the era of relatively modest growth before 1980, disparities in growth were also muted. But when growth took off, it was highly unequal across the country, aggravating inequalities in standards of living, especially between the peninsular states and the hinterland and northeastern states (Kochhar et al. 2006). In Aravind Adiga's (2008) Booker Prize-winning novel, *The White Tiger*, the protagonist describes the geography-based explanation of India's development thus: "Please understand, Your Excellency that India is two countries: an India of Light, and an India of Darkness. The ocean brings

[7]The *World Inequality Report* compares the distributional implications of the structural transformation paths chosen by India and China and how the paths followed by the two countries are mirrored in the evidence on the distribution of personal income (Alvaredo et al. 2018).

Figure 4
Income Convergence/Divergence: India, China, and the World



Source: Per capita GDP for Indian states comes from Handbook of India Statistics by the Reserve Bank of India. Since no data for state level PPP is available, the state level per capita GDP in rupee is normalized with India per capita GDP in PPP US$/India per capita GDP in rupee. GDP for Chinese provinces is taken from the National Bureau of Statistics of China, these numbers are also converted into PPP US$. For all other countries, the Maddison dataset is used.
Note: The figure is a scatter plot of average growth rate for Indian states, Chinese provinces, and nation states for two ten-year periods, 1994–2004 and 2004–2014 (*y*-axis), against per capita GDP in the start of the respective ten-year period, 1994 and 2004 (*x*-axis).

light to my country. Every place on the map of India near the ocean is well off. But the river brings darkness to India."[8]

The natural framework for assessing this theory is a simple test for convergence (Barro and Sala-i Martin 1992).[9] Consider the per capita GDP across the states of India, and then look at the average per capita GDP growth rate for these states. If regions with lower per capita income grow faster on average, convergence occurs; conversely, if those with higher per capita income grow faster, divergence results.

Figure 4 plots the initial per capita GDP against the average growth rate over the next ten years for three different categories: states in India, provinces in China, and all the countries of the world, with Figure 4, panel A showing the period 1994–2004 and Figure 4, panel B the period from 2004–2014. In the first period, provinces in

[8]The considerable heterogeneity of growth outcomes makes India a crucible for illustrating and understanding the many patterns and theories of economic development: Punjab and Haryana were centers of the boom in agriculture during the Green Revolution; Gujarat and Maharashtra are manufacturing successes; Karnataka and Tamil Nadu and many cities across the country have fueled growth through skilled services; Kerala's growth owes to remittances from its large export of labor to the Middle East; central and eastern India exhibit many of the pathologies associated with the natural resources curse; and poorer states as well as those in the northeast are susceptible to an aid curse (Ministry of Finance 2017, chapter 13).
[9]Technically, this is called beta-convergence, which is distinct from sigma-convergence; the latter refers to a decline in the dispersion of real per capita income (Quah 1996).

China do not show a strong trend and neither do countries of the world, but states in India show clear signs of divergence (upward-sloping blue line). In the second period, provinces in China and countries of the world start converging (downward-sloping red and green lines), but states in India now start diverging more rapidly.

Therefore, the evidence so far suggests that in India, regional/spatial catch-up remains elusive. The striking contrast between the results in India versus those in China and internationally poses an important puzzle. If a state/country is capital-scarce and poor, then it seems as if returns to capital should be high and the area should be able to attract capital and technology, thereby raising its productivity and enabling catch-up with richer states/countries. Within India, where borders are porous, this process of convergence has failed. But across countries where borders are much thicker (because of restrictions on trade, capital, labor, and technology), convergence has occurred. That pattern is not easy to explain.

One possible explanation is that convergence fails to occur because of traps relating to governance and state capacity. Poor governance could make the risk-adjusted returns on capital low, even in capital-scarce states. Moreover, greater labor mobility or exit from these areas, especially of the higher skilled, could further worsen governance, creating a vicious cycle. Another possible explanation relates back to India's structural pattern of growth. If growth has been skill-intensive, there is no reason why labor productivity would necessarily be high in capital-scarce states. Unless the less developed regions are able to generate skills (in addition to good governance), convergence may not occur.

Chauvin et al. (2016) argue that India is both underurbanized and has too few large cities (violating Zipf's law). So, if India is still realizing agglomeration economies from early urbanization, larger, richer regions will benefit at the expense of smaller, poorer ones. Somewhat unexpectedly, the Indian pattern of divergence is coming to resemble the more recent emergence of divergence amongst the cities and regions in the United States and the nations of Europe and for similar reasons of a rising importance of agglomeration and skills (Krugman 1991; Hendrickson, Muro, and Galston 2018; Redding and Rossi-Hansberg 2017). A skill-based technical bias in labor demand emerges, and its persistence manifests in uneven growth (Card and DiNardo 2002; Giannone 2019).

Caste and Religion

Caste and religion are distinctive markers of Indian society, a perennial source of cleavage and conflict before and after independence. (A famous quip is that "Indians don't cast their votes, they vote their castes.")

The five largest social groups in India are Hindu Upper Castes, the historically most privileged category, together with "Scheduled Castes," "Scheduled Tribes," "Other Backwards Castes," and Muslims. India's constitution granted special status for the so-called Scheduled Castes that were deemed "untouchable" because of being outside the caste hierarchy. Such status was also granted for a group called Scheduled Tribes, indigenous inhabitants of regions in central and eastern India as well as in the northeast. This status took the form of guaranteed minimum political

representation, admission to public sector educational institutions, and employment in the public sector. This special status was extended in the 1990s to Other Backwards Castes. About 80 percent of India's population is Hindu, 15 percent is Muslim, with the rest being a mixture of Christian, Sikh, and other religious groups. However, no special status has been awarded to minority religious groups, like Muslims.

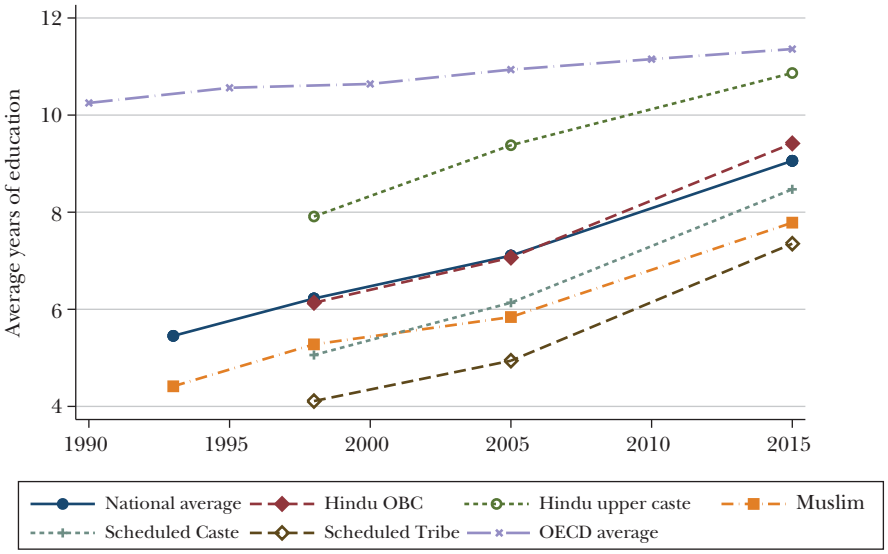
On one side, economic growth could help to mitigate these traditional divisions. This hope is evoked by Suketu Mehta's (2005) description in the book *Maximum City* of a hand extending to help a passenger get in the local train in Mumbai: "And at the moment of contact, they do not know if the hand that is reaching for theirs belongs to a Hindu or Muslim or Christian or Brahmin or untouchable or whether you were born in this city or arrived only this morning. All they know is that you're trying to get to the city of gold, and that's enough. Come on board, they say. We'll adjust."

On the other side, economic growth also interacts with traditional cleavages of caste and religion to reinforce existing hierarchies. The dominance of Hindu upper castes in private sector jobs, academia, and civil services is a case in point. Electoral politics along caste lines though has helped to some extent in pushing resources towards the marginalized. What has been the culmination of these competing forces in terms of measurable outcomes for the various communities?

There is evidence that economic growth has indeed played an important part in diminishing cleavages of caste and religion, especially in urban India. For example, Varshney (2002) analyzes three pairs of cities with a history of Hindu-Muslim violence and argues that civic engagement, such as integrated business organizations, trade unions, political parties, and professional associations, are able to control outbreaks of ethnic violence. Kapur et al. (2010) surveyed Dalit (untouchable) households in Uttar Pradesh (a poor state) and found enormous changes in social norms between 1990 and 2010. A combination of economic growth, migration, and the acquisition of political power has meant that previous social taboos on co-dining and co-mingling as well as rigid caste occupation links were breaking down. Relatedly, Kapur et al. (2014) document the stories of a number of "Dalit entrepreneurs" who have been able to build midsized businesses.

Using the National Family Health Surveys, which have so far conducted four rounds—1992, 1998, 2005, and 2015—we assess progress on the key indicator of education (similar analysis can be conducted for wealth and height). Figure 5 plots educational outcomes of 15–29 year-olds, measured as average years of schooling for the five largest social groups. There are two clear findings. At least in terms of educational quantities (not necessarily in terms of quality), India's most privileged groups are converging to the global frontier. Within India, however, there is more limited convergence. The gap with the Hindu Upper Castes has shrunk somewhat for the Other Backwards Castes (from 1.8 years in 1998 to 1.4 years in 2015), for the Scheduled Castes (from 2.9 years in 1998 to 2.4 years in 2015), and Scheduled Tribes (from 3.8 years in 1998 to 3.5 years in 2015), but has widened for the Muslims (from 2.6 years in 1998 to 3.1 years in 2015). There is catch-up, but it is slow for many groups

Figure 5
Educational Attainment of Age Group 15–29 across Social and Religious Groups, 1992–2015



Source: For each social group in India, the data source is four rounds of National Family Health Survey (all four rounds), which can be obtained by applying for the Demographic and Health Survey. For the OECD countries, the Barro-Lee dataset is used. Since the Barro-Lee dataset ends in 2010, we extrapolate it to 2015 using the average slopes previously.
Note: The figure plots the average years of education of adults aged 15–29 years for various social groups (*y*-axis) over time (*x*-axis).

and absent for Muslims.[10] The broad pattern of continuing inequality in educational outcomes across social and religious groups is consistent with theories that emphasize unequal access to learning as an instrument of elite dominance (for commentary on India, see Weiner 1997; for a general argument, see Fukuyama 2011).

Relatedly, using various data sources, Asher, Novosad, and Rafkin (2018) document rising intergenerational mobility for Scheduled Castes and declining mobility for Muslims. Since the two population sizes are approximately the same, these two effects cancel each other to produce almost no intergenerational mobility in the aggregate over the last few decades. Banerjee, Gethin, and Piketty (2019) analyze electoral data to conclude that the traditional cleavages may actually be on the rise.

[10] In 2009, India adopted the Right to Education wherein every child was granted the fundamental right to (free) education. Its impact on educational attainment is not yet well documented, but India has attained universal enrollment in primary education.

Gender and Children

Although some part of the reduction in gender inequality can be explained by the process of economic development, society-specific factors play a big role (Jayachandran 2015). In India, in particular, a number of cultural factors may cause gender development to lag growth dynamism. Examples include patrilocality (women moving after marriage to live with the husband's parents), patrilineality (titles and property passing on to sons), rituals performed by oldest sons, dowry system, old-age support provided by sons, and strong notions of cultural purity of women.

India has made progress on a number of gender-related measurables. The 2018 *Economic Survey of India* (Ministry of Finance 2018, chapter 7) showed improvements in 14 out of 17 indicators, relating to agency, attitudes, and outcomes. For example, India's score improved on agency for women in decision-making regarding household purchases and visiting family and relatives, on the experience of physical and sexual violence, and on educational attainment.

But two other striking outcomes paint a disappointing picture: contraception and female labor force participation. Nearly 47 percent of Indian women do not use any contraception, and of those who do, less than one-third use female-controlled reversible contraception. In 2015, India was an outlier by more than 50 percentage points in the use of sterilization for women as means of contraception amongst a group of low- and middle-income countries.

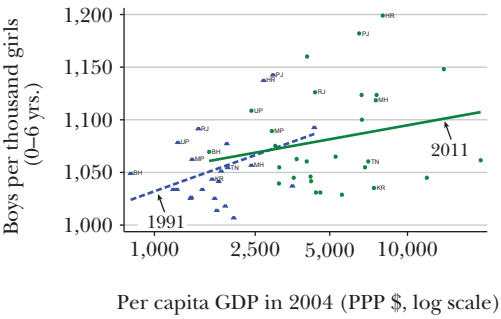
Female labor force participation in India has been declining from about 35 percent in 1990 to about 28 percent in 2015. For perspective, the female labor force participation rate in Indonesia in 2015 was almost 50 percent; in China, it was above 60 percent. In addition, the gap between India's labor force participation rate and the rate of countries with similar per capita GDP is widening, not narrowing.

India's gender problems are perhaps steeped in a deeper form of discrimination—a strong preference for male children, documented in Figure 6. A malign version of this preference, facilitated by the now-banned ultrasound technology, involves selective sex abortion and female foeticide (Sen 1990; Anderson and Ray 2010). India's sex ratio at birth increased from 1,060 boys born for every 1,000 girls in 1970 to 1,106 in 2014, widening its gap from the biological norm of 1,050. The usual pattern around the world is that countries with higher income levels have sex ratios at birth closer to the expected biological norm (Jayachandran 2015). But within India, states with a higher per capita GDP tend to have more unbalanced sex ratios at birth, and this perverse relationship has not changed between 1991 and 2011 (as shown in Figure 6, panel A). This suggests ominously that future economic growth may not necessarily reduce this imbalance.

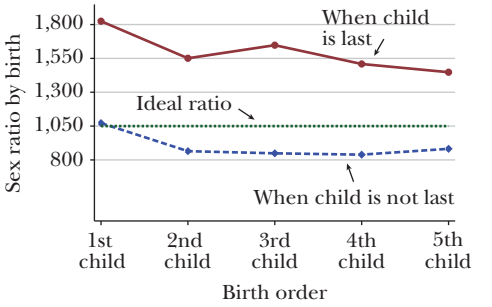
The less malign but no less important version of son preference relates to fertility choices. Even without selective sex abortion and female foeticide, parents may choose to keep having children until they get the desired number of sons. Such a meta-preference for a son manifests itself in sex ratios depending on birth order (shown in Figure 6, panel B). If the child is not the last (lower dotted line), the sex ratio is skewed in favor of girls (850 boys per 1,000 girls) and below the ideal sex

Figure 6
"Missing Women" and "Unwanted Girls"

A: Gender ratio and income per capita across states of India, 1991 and 2011



B: Sex ratio by birth order, 2015



Source: The per capita GDP for each state is calculated in the same way as Figure 4. The data on sex ratio for 1991 and 2011 comes from the Census of India. The sex ratio at birth given birth order comes from the National Family Health Survey.
Note: Panel A plots the number of boys per thousand girls aged 0–6 years (*y*-axis) against per capita GDP (*x*-axis). Panel B plots the sex ratio (*y*-axis) against the order of birth (*x*-axis).

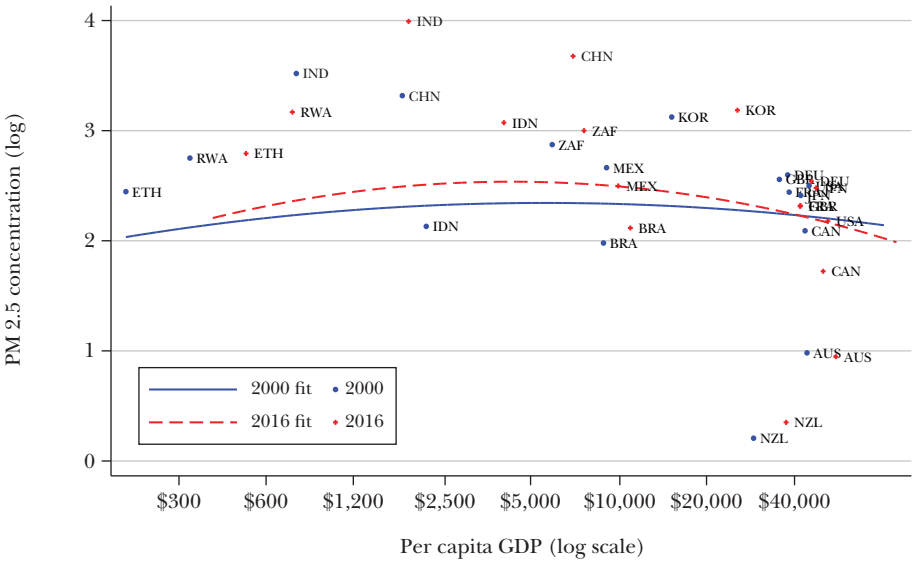
ratio (horizontal line). In contrast, if the child (at any birth order) is the last child (top line), the sex ratio is massively skewed in favor of boys (1,500 to 1,000).

We can quantify both forms of son preference. Because the sex ratio at birth is so skewed, about 40 million women are "missing," to use Sen's (1990) famously evocative term. Similarly, because Indians seem to stop having children as soon as a son is born, as reflected in sex ratio as a function of the gender of the last-born child, over 20 million "unwanted girls" are born in India (Ministry of Finance 2018). The twin phenomena of missing women and unwanted girls—malign and meta-son preference, respectively—reflect Indian society's deepest gender discrimination.

This bias in favor of sons, especially the eldest one, can be detrimental to the resources available to other children and to female children in particular. Jayachandran and Pande (2017) show how favoring resources devoted to eldest sons reduces the investments in other children, especially girls, contributing to the phenomenon of "stunting," which refers to a situation in which malnourishment leads to children whose height-for-age is two standard deviations below the medians calculated by the World Health Organization. Gender discrimination begets child neglect.

In addition to the preference for the first-born male, another main cause of stunting is that a large majority of people—especially in the rural areas—defecate in the open, which leads to diarrhea and less absorption of nutrients amongst children (Coffey and Spears 2017). According to the Demographic and Health Survey, 52 percent of children in India were stunted in 1998, and although the number dropped to 38 percent in 2015, India remains a distinct outlier in the extent of stunting for its level of per capita income.

Figure 7
Pollution and Development, 2000 and 2016



Source: The PM 2.5 concentration comes from the WDI.
Note: The figure is a scatter plot of the PM 2.5 concentration of nation states (*y*-axis) against their per capita GDP (*x*-axis).

Environment

As measured by particulate matter of 2.5 microns or more in the air, the so-called PM 2.5 index, 22 of the top 30 most polluted cities in the world are in India (according to the World Health Organization Global Ambient Air Quality Database as of 2018). Greenstone et al. (2015) estimate that around 660 million people, over half of India's population, live in areas that exceed the Indian National Ambient Air Quality Standard for fine particulate pollution. In 2017 alone, 1.24 million deaths (12.4 percent of all deaths) in India were attributable to air pollution (India State-Level Disease Burden Initiative Air Pollution Collaborators 2019).

The environmental "Kuznets curve" suggests that environmental quality may first decline and then rise with capita GDP (Shafik and Bandyopadhyay 1992; Grossman and Krueger 1995; for an overview in this journal, see Dasgupta et al. 2002). Although the theory is controversial, the intuition is that in the initial stages of development, growth will lead to greater output and consumption, and especially if accompanied by a move toward energy-intensive manufacturing, also to greater pollution. But at some point, a combination of consumer preference for a better environment, a shift toward less resource-intensive services, and the availability of greener technology should lead to a positive impact of growth on environmental quality.

Figure 7 plots the population-weighted PM 2.5 index against per capita GDP for the broad cross section of countries for 2000 and 2016. It is hard to detect any

pattern in the data, let alone a Kuznets curve relationship. But what is unmistakable is that India (along with China) is a striking outlier and would be so even if there were a U-shaped relationship. In addition, India has over time become more of an outlier both compared to the average country (represented by the two lines of best fit) and even compared to China. Levels of pollution in India have risen sharply, more so than should be warranted by economic activity, despite services and not manufacturing being the primary driver of growth. Such high levels of pollution reflect weak regulation and enforcement and are symptomatic of weak state capacity.

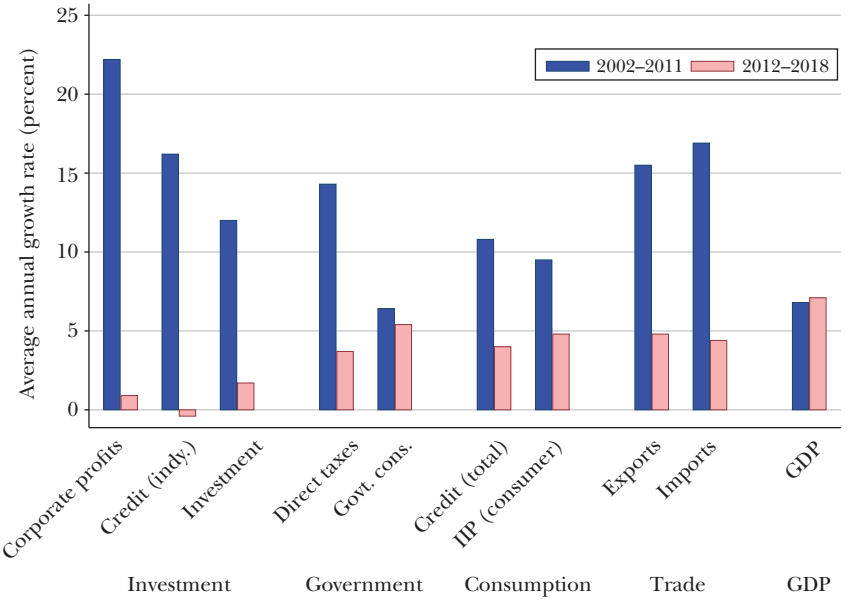
India faces other environmental problems as well. For example, rapid urbanization and indiscriminate use of water for irrigation purposes in agriculture have created a severe groundwater problem over the last two decades. A complex web of input and output subsidies in farming lead to overexploitation of groundwater (Badiani-Magnusson and Jessoe 2019; Chatterjee, Lamba, and Zaveri 2017). Groundwater levels have dropped from 8 meters below ground level to 16 meters below ground level in northwestern India and from 1 to 8 meters below ground level in the rest of the country. For perspective, the groundwater loss in India is orders of magnitude larger than water depletion in California's Central Valley during the same period (Zaveri et al. 2016). Globally speaking, the problem of accelerated groundwater depletion is the most severe in South Asia in general, and in India in particular (Aeschbach-Hertig and Gleeson 2012, figure 2).

Data Issues: Is India's Recent Dynamism Overstated?

One possible reconciliation of the development-lagging-dynamism hypothesis could be an overstatement of India's dynamism. Beginning in 2011–2012, as part of periodic revisions, India's government introduced a number of changes in estimating its National Income Accounts. For example, the base year was changed from 2004–2005 to 2011–2012, and the data sources were comprehensively expanded to use financial accounts of over 600,000 companies. As a result, calculations moved from predominantly volume-based estimates of gross value added to value-based estimates that potentially better capture economic changes in a modern, dynamic economy. A robust debate has since ensued on India's growth numbers (Bhattacharya 2019; Dholakia, Nagaraj, and Pandya 2018; Nagaraj and Srinivasan 2017; Ministry of Finance 2015, 2017).

Because India measures GDP from the production side, a natural question to ask is whether these production-side estimates can be validated by demand-side indicators such as investment, exports, imports, and credit, which are measured independently and arguably more reliable (Subramanian 2019a,b). In the decade preceding the methodological changes, India's measured GDP growth exhibits a strong correlation with other demand indicators: GDP growth of about 7.5 percent was accompanied by double-digit growth in investment (13 percent) and exports (15 percent), which are critical drivers of medium-term growth. But since 2011, the Indian economy experienced a series of shocks: exports declined after the global

Figure 8
Is India's Recent GDP Growth Overstated?



Source: Real investment, credit, and government consumption are obtained by deflating nominal values by the Consumer Price Inflation index. All these numbers and GDP growth are taken from WDI. Real exports and imports are also taken from WDI. Corporate profits come from the Prowess database. Credit to industry comes from the Reserve Bank of India's database of the Indian economy. The Index of Industrial Production (consumer goods) comes from the Ministry of Statistics and Programme Implementation, Government of India.
Note: The figure plots ten variables averaged over two different time periods: 2002–2011 and 2012–2018.

financial crisis; the "twin balance sheet" crises stifled credit and investment as described earlier; the "taper tantrum" affected macroeconomic stability; two successive agricultural droughts diminished rural demand; and a demonetization hit the informal labor market. During this time, the main positive shock was an improvement in India's terms of trade as oil prices declined.

Figure 8 plots a series of demand-side indicators (investment, exports, government consumption, and private consumption) and associated proxies, before and after the changed GDP measurement. The annual average growth of all these indicators declined by between 10 and 20 percentage points. For example, investment fell from an annual rate of 12 to 1.7 percent; credit to industry from 16.2 to –0.4 percent; exports from 15.5 to 4.8 percent; and perhaps most tellingly, imports from 16.9 to 4.4 percent. Despite these large declines in every component of demand, GDP growth, as measured by the revised production-side methodology, actually increased—which seems implausible.

Other pieces of evidence reinforce the puzzle. Comparisons to other emerging economies in the same time frame show that many of these countries experienced lesser shocks, but still saw a larger decline in growth. Also, correlations between different measures of manufacturing move together before the measurement change but decouple thereafter. Further, an estimation of India's wedge between the GDP deflator and consumer price index shows a very small discrepancy pre-2011 and large one (in fact the second largest in a sample of comparison countries) post-2011. The GDP deflator anomaly is consistent in timing, sign, and magnitude with real GDP growth anomaly (Subramanian 2019b).

In summary, data issues call into question India's growth dynamism of the current decade. But how seriously would these issues affect the underlying narrative of long-run economic growth? Suppose that the magnitude of overestimation is 2.5 percent per year since 2011 (Subramanian 2019a). In this case, India's 38-year annual average per capita growth rate would decline from 4.6 percent to 4.2 percent, which would still be exceptional performance and in fact would preserve India in the list of the top ten fastest and stable growing economies. Thus, the basic narrative of dynamism with incommensurate development would remain valid.

Looking Ahead: The Challenges of Development with Dynamism

The core argument of this paper is that although India has experienced rapid and stable economic growth for nearly four decades, the resulting development has been limited on a number of dimensions, including structural change, regional divergence, inadequate convergence across caste and religion, and underperformance on issues related to gender and children, and environmental outcomes. This pattern raises two obvious questions: How can the dynamism be sustained going forward? How can the concomitant development transformations be accelerated?

The sustainability of growth—which in late 2019 has cratered to a near-standstill—will be determined by structural factors salient amongst which is the "twin balance sheet challenge" initiated by the toxic legacy of the credit boom of the 2000s. Recently, the rot of stressed loans has spread from the public sector banks to the nonbank financial sector, and on the real side, from infrastructure companies to most notably the real estate sector with the latter threatening middle class savings. This contagion owes both to overall weak economic growth and slow progress in cleaning up bank and corporate balance sheets. A failure to resolve this challenge could mean a reprisal of the Japanese experience of nearly two decades of lost growth, but at a much lower level of per capita income. India's development experience could end up being a transition from socialism without entry to capitalism without exit because weak regulatory capacity and lack of social buy-in will have impeded the necessary creative destruction.

Over the longer run, India's unusual structural transformation will pose severe challenges. A high share of India's workforce is employed in low-productivity occupations—agriculture and informal manufacturing. For creating the jobs of the future, India can either try to rehabilitate the unskilled manufacturing sector or try to lay a groundwork for sustaining a more skill-intensive pattern of growth. Attempting the former would be a history-defying achievement because there are not many examples of durable reversals of premature deindustrialization. At minimum, this approach would involve enormous construction of infrastructure along with reforming the panoply of laws and regulations that disincentivize both firm expansion and exit. Moreover, the new international environment, especially the backlash against globalization and labor-saving technology, will make it difficult for India to sustain a policy of export-led growth based on low-skill manufacturing.

On the other hand, sustaining a skill-intensive pattern with a greater focus on education (and skills development) poses its own difficulties. This approach carries the risk that one or two generations of those who are currently unskilled will be left out. Another problem is that India's performance at building skills has been unsatisfactory. Learning outcomes in primary education are poor and stagnant, despite years of rapid economic growth which has increased the private returns to education (Muralidharan and Singh 2019). For example, the 2018 Annual Status of Education Report (ASER Center 2018) revealed that less than 30 percent of students in grade three were able to solve problems of reading and writing at the level of grade two, and less than 30 percent of students in grade five were able to do math problems associated with grade two. These learning gaps are high and rising (Ministry of Finance 2018, chapter 5). For India, building a skill-intensive model of growth on such tenuous foundations will be difficult. Why there has not been greater political salience for improving education is one of the deeper puzzles about Indian politics, deserving of extensive research.

One long-run perspective for dynamism and development in India relates back to the idea of political institutions as a cause for high and equitable growth (for example, North 1990; Acemoglu and Robinson 2012). India has long been a country where the institutions of democratic governance have been much more advanced than other aspects of development: a plot of average income over time against the strength of political institutions shows India to be an extreme outlier. The hopeful way to interpret this exceptionalism is to argue that the strength of India's democratic institutions can provide a basis for dramatic improvements in development: that is, India has been an economic underperformer relative to its political development, and there is considerable scope for mean reversion.

The pessimistic interpretation is that India's adoption of democracy at such an early stage of development may have created a situation of weak state capacity, which has become hardwired into the Indian development model (for discussion, see Mehta 2003 and the article by Kapur in this symposium): India being an outlier on the politics-economics relationship is then a feature, not a bug. A plausible mechanism is the following: a cleavaged, precocious democracy created early

pressures to redistribute and in inefficient ways. As a result, the Indian state never acquired the legitimacy stemming from effective provision of public goods such as health and education. Moreover, identity politics in a cleavaged society meant that democracy created greater pressures for excludable club goods rather than broader public goods. This engendered distrust and elicited exit, especially by the middle class, depriving the state of resources and further worsening state capacity.

The somber conclusion of this line of analysis is that Indian underperformance in broader categories of development is not an aberration that time will necessarily correct. India cannot afford to be complacent that its robust democracy will ensure economic dynamism and broader development. Restoring dynamism and accelerating structural transformation will require the industrious political work, in Weber's (1919) famous phrase, "of a strong and slow boring of hard boards."

■ *This paper draws upon ideas developed in the* Economic Surveys of India, *written at India's Ministry of Finance when Subramanian served as Chief Economic Adviser and Lamba was a member of the team. The authors are extremely grateful to the entire team that worked on the Economic Surveys; to Josh Felman and Devesh Kapur for valuable inputs; to Abhishek Anand, Kapil Patidar, Sagar Saxena, and especially Abhishek Rai for excellent research assistance; and to Gordon Hanson, Heidi Williams, and especially Timothy Taylor for superb substantive and editorial comments.*

References

Acemoglu, Daron, and James A. Robinson. 2012. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York: Crown, Crown Business.

Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared. 2009. "Reevaluating the Modernization Hypothesis." *Journal of Monetary Economics* 56 (8): 1043–58.

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A. Robinson. 2019. "Democracy Does Cause Growth." *Journal of Political Economy* 127 (1): 47–100.

Adelman, Jeremy. 2013. *Worldly Philosopher: The Odyssey of Albert O. Hirschman.* Princeton: Princeton University Press.

Adhia, Hasmukh, and Arvind Subramanian. 2016. "One India, One Market." *The Hindu*, July 19. https://www.thehindu.com/opinion/lead/One-India-one-market/article14496026.ece.

Adiga, Aravind. 2008. *The White Tiger.* New York: Free Press.

Aeschbach-Hertig, Werner, and Tom Gleeson. 2012. "Regional Strategies for the Accelerating Global Problem of Groundwater Depletion." *Nature Geoscience* 5: 853–61.

Ahluwalia, Montek S. 2002. "Economic Reforms in India since 1991: Has Gradualism Worked?" *Journal of Economic Perspectives* 16 (3): 67–88.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. 2003. "Fractionalization." *Journal of Economic Growth* 8 (2): 155–94.

Alvaredo, Facundo, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2018. *World*

Inequality Report. Paris: World Inequality Lab.

Amirapu, Amrit, and Arvind Subramanian. 2015. "Manufacturing or Services? An Indian Illustration of a Development Dilemma." Center for Global Development Working Paper 357.

Anderson, Siwan, and Debraj Ray. 2010. "Missing Women: Age and Disease." *Review of Economic Studies* 77 (4): 1262–1300.

ASER Center. 2018. *Annual Status of Education Report*. New Delhi: ASER Center.

Asher, Sam, Paul Novosad, and Charlie Rafkin. 2018. "Intergenerational Mobility in India: Estimates from New Methods and Administrative Data." https://www.dartmouth.edu/~novosad/anr-india-mobility.pdf.

Badiani-Magnusson, Reena, and Katrina Jessoe. 2019. "Electricity Prices, Groundwater, and Agriculture: The Environmental and Agricultural Impacts of Electricity Subsidies in India." In *Agricultural Productivity and Producer Behavior*, edited by Wolfram Schlenker, 157–83. Chicago: University Chicago Press.

Banerjee, Abjijit, Amory Gethin, and Thomas Piketty. 2019. "Evidence from the Changing Structure of Electorates, 1962–2014: Growing Cleavages in India?" *Economic and Political Weekly* 54 (11): 34–44.

Banerjee, Abhijit, and Rohini Somanathan. 2007. "The Political Economy of Public Goods: Some Evidence from India." *Journal of Development Economics* 82 (2): 287–314.

Bardhan, Pranab. 1999. *The Political Economy of Development in India*. London: Oxford University Press.

Barro, Robert J., and Xavier Sala-i Martin. 1992. "Convergence." *Journal of Political Economy* 100 (2): 223–51.

Barro, Robert J., and Jong Wha Lee. 2013. "A New Data Set of Educational Attainment in the World, 1950–2010." *Journal of Development Economics* 104: 184–98.

Basu, Kaushik. 2018. "A Short History of India's Economy: A Chapter in the Asian Drama." United Nations World Institute for Development Economics Research Working Paper 2018/124.

Bhagwati, J. N., and P. Desai. 1970. *India: Planning for Industrialization*. London: Oxford University Press.

Bhagwati, J. N., and T. N. Srinivasan. 1995. *India's Economic Reforms*. New Delhi: India Ministry of Finance.

Bhandari, Pranjul, and Rohit Lamba. 2016. "25 Years of 1991 Reforms: In Praise of Creative Destruction." *Open*, June 25. https://openthemagazine.com/features/business/25-years-of-1991-reforms-in-praise-of-creative-destruction/.

Bhattacharya, Pramit. 2019. "New GPD Series Faces Fresh Questions after NSSO Discovers Holes." *Livemint*, May 10. https://www.livemint.com/news/india/new-gdp-series-faces-fresh-questions-after-nsso-discovers-holes-1557250830351.html.

Bosworth, Barry, and Susan M. Collins. 2008. "Accounting for Growth: Comparing China and India." *Journal of Economic Perspectives* 22 (1): 45–66.

Card, David, and John E. DiNardo. 2002. "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles." *Journal of Labor Economics* 20 (4): 733–83.

Center for Systemic Peace. 2015. "Polity IV Dataset." https://www.systemicpeace.org/polityproject.html.

Centre for Monitoring Indian Economy. 2018. "ProwessIQ Database." https://prowessiq.cmie.com/.

Chatterjee, Shoumitro, Rohit Lamba, and Esha Zaveri. 2017. "The Water Gap: Environmental Effects of Agricultural Subsidies in India." https://pdfs.semanticscholar.org/ec94/3694c643bad9be4b5b0332ae71c2a76f5197.pdf.

Chauvin, Juan Pablo, Edward Glaeser, Yueran Ma, and Kristina Tobio. 2016. "What Is Different about Urbanization in Rich and Poor Countries? Cities in Brazil, China, India and the United States." National Bureau of Economic Research Working Paper 22002.

Chodorow-Reich, Gabriel, Gita Gopinath, Prachi Mishra, and Abhinav Narayanan. 2018. "Cash and the Economy: Evidence from India's Demonetization." National Bureau of Economic Research Working Paper 25370.

Clark, Colin. 1940. *The Conditions of Economic Progress*. London: MacMillan.

Coffey, Diane, and Dean Spears. 2017. *Where India Goes: Abandoned Toilets, Stunted Development, and the Costs of Caste*. Noida: HarperCollins India.

Crabtree, James. 2018. *The Billionaire Raj: A Journey through India's New Gilded Age*. New York: Tim Duggan Books.

Das, Gurcharan. 2012. *India Grows at Night: A Liberal Case for a Strong State*. New Delhi: Penguin Random House India.

Dasgupta, Susmita, Benoit Laplante, Hua Wang, and David Wheeler. 2002. "Confronting the Environmental Kuznets Curve." *Journal of Economic Perspectives* 16 (1):147–68.

DeLong, J. Bradford. 2003. "India since Independence: An Analytic Growth Narrative." In *In Search*

of Prosperity: Analytic Narratives on Economic Growth, edited by Dani Rodrik, 184–204. Princeton: Princeton University Press.

de Mesquita, Bruce Bueno, and George W. Downs. 2005. "Development and Democracy: Richer but Not Freer." *Foreign Affairs*, September 1. https://www.foreignaffairs.com/articles/2005-09-01/development-and-democracy.

Demographic and Health Surveys Program (DHS Program). 2019. *Demographic and Health Surveys.* https://dhsprogram.com/.

Dholakia, Ravindra H., R. Nagaraj, and Manish Pandya. 2018. "Manufacturing Output in New GDP Series." *Economic and Political Weekly* 53 (35).

Drèze, Jean, and Amartya Sen. 2013. *An Uncertain Glory: India and Its Contradictions.* Princeton: Princeton University Press.

Engerman, Stanley L., and Kenneth L. Sokoloff. 2005. "The Evolution of Suffrage Institutions in the New World." *Journal of Economic History* 65 (4): 891–921.

Fukuyama, Francis. 1989. "The End of History?" *National Interest* 16: 3–18.

Fukuyama, Francis. 2011. *The Origins of Political Order: From Prehuman Times to the French Revolution.* New York: Farrar, Straus and Giroux.

George, Siddharth, and Arvind Subramanian. 2015. "Transforming the Fight against Poverty in India." *New York Times*, July 22. https://www.nytimes.com/2015/07/23/opinion/transforming-the-fight-against-poverty-in-india.html.

Giannone, Elisa. 2019. "Skill-Biased Technical Change and Regional Convergence." https://drive.google.com/file/d/1dDBJeI-WXDbZ95DfpRGM21EPBPrm_8Si/view.

Greenstone, Michael, Janhavi Nilekani, Rohini Pande, Nicholas Ryan, Anant Sudarshan, and Anish Sugathan. 2015. "Lower Pollution, Longer Lives: Life Expectancy Gains if India Reduced Particulate Matter Pollution." *Economic and Political Weekly Special Article* 50 (8): 40–46.

Groningen Growth and Development Centre. 2015. "The GGDC 10-Sector Database." https://www.rug.nl/ggdc/productivity/10-sector/.

Grossman, Gene M., and Alan B. Krueger. 1995. "Economic Growth and the Environment." *Quarterly Journal of Economics* 110 (2): 353–77.

Guha, Ramachandra. 2007. *India after Gandhi: The History of the World's Largest Democracy.* New York: Harper Perennial.

Hendrickson, Clara, Mark Muro, and William A. Galston. 2018. *Countering the Geography of Discontent: Strategies for Left-Behind Places.* Washington, DC: Brookings Institute.

Hensel, Paul R. 2018. "The ICOW Colonial History Data Set, Version 1.1." https://www.paulhensel.org/icowcol.html.

Herrendorf, Berthold, Richard Rogerson, and Ákos Valentinyi. 2014. "Growth and Structural Transformation." In *Handbook of Economic Growth*, Vol. 2, edited by Philippe Aghion and Steven N. Durlauf, 855–941. Amsterdam: North-Holland.

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

Huntington, Samuel P. 1969. *Political Order in Changing Societies.* Cambridge, MA: Yale University Press.

India State-Level Disease Burden Initiative Air Pollution Collaborators. 2019. "The Impact of Air Pollution on Deaths, Disease Burden, and Life Expectancy across the States of India: The Global Burden of Disease Study 2017." *Lancet Planetary Health* 3 (1): e26–39.

Jayachandran, Seema. 2015. "The Roots of Gender Inequality in Developing Countries." *Annual Review of Economics* 7: 63–88.

Jayachandran, Seema, and Rohini Pande. 2017. "Why Are Indian Children So Short? The Role of Birth Order and Son Preference." *American Economic Review* 107 (9): 2600–2629.

Joshi, V., and I. M. D. Little. 1996. *India's Economic Reforms, 1991–2001.* London: Oxford University Press.

Kapur, Devesh, Chandra Bhan Prasad, Lant Pritchett, and D. Shyam Babu. 2010. "Rethinking Inequality: Dalits in Uttar Pradesh in the Market Reform Era." *Economic and Political Weekly* 45 (35): 39–49.

Kapur, Devesh, D. Shyam Babu, and Chandra Bhan Prasad. 2014. *Defying the Odds: The Rise of Dalit Entrepreneurs.* New Delhi: Random House India.

Kochhar, Kalpana, Utsav Kumar, Raghuram Rajan, Arvind Subramanian, and Ioannis Tokatlidis. 2006. "India's Pattern of Development: What Happened, What Follows?" *Journal of Monetary Economics* 53 (5): 981–1019.

Kohli, Atul. 2010. *Democracy and Development in India: From Socialism to Pro-Business.* London: Oxford University Press.

Krugman, Paul. 1991. "Increasing Returns and Economic Geography." *Journal of Political Economy* 99 (3): 483–99.

Kuznets, Simon. 1957. "Quantitative Aspects of the Economic Growth of Nations: II. Industrial Distribution of National Product and Labor Force." *Economic Development and Cultural Change* 5 (4): 1–111.

Lewis, W. Arthur. 1954. "Economic Development with Unlimited Supplies of Labour." *Manchester School* 22 (2): 139–91.

Lipset, Seymour Martin. 1959. "Some Social Requisites of Democracy: Economic Development and Political Legitimacy." *American Political Science Review* 53 (1): 69–105.

Maddison Project. 2018. *Maddison Historical Statistics*. https://www.rug.nl/ggdc/historicaldevelopment/maddison/.

McMillan, Margaret, Dani Rodrik, and Íñigo Verduzco-Gallo. 2014. "Globalization, Structural Change, and Productivity Growth, with an Update on Africa." *World Development* 63 (4): 11–32.

Mehta, Pratap Bhanu. 2003. *The Burden of Democracy*. New York: Penguin.

Mehta, Suketu. 2005. *Maximum City: Bombay Lost and Found*. New York: Vintage Books.

Ministry of Finance. 2015. *Economic Survey of India 2014–15*. Delhi: Oxford University Press India.

Ministry of Finance. 2017. *Economic Survey of India 2016–17*. Delhi: Oxford University Press India.

Ministry of Finance. 2018. *Economic Survey of India 2017–18*. Delhi: Oxford University Press India.

Ministry of Statistics and Programme Implementation. 2018. "Index of Industrial Production." http://mospi.nic.in/iip.

Mohan, Rakesh. 2018. *India Transformed: 25 Years of Economic Reforms*. Washington, DC: Brookings Institution.

Muralidharan, Karthik, and Abhijeet Singh. 2019. "Learning Levels Will Not Improve by Spending More on Education." *Hindustan Times*, April 17th. https://www.hindustantimes.com/columns/learning-levels-will-not-improve-by-spending-more-on-education/story-Ej6tQgP6f9JOTQi6HuvniO.html.

Nagaraj, R., and T. N. Srinivasan. 2017. "Measuring India's GDP Growth: Unpacking the Analytics and Data Issues behind a Controversy That Has Refuses to Go Away." In *India Policy Forum 2016–17*, Vol. 13, edited by Shekhar Shah, Barry Bosworth, and Karthik Muralidharan, 73–128. New Delhi: SAGE Publications.

National Bureau of Statistics of China. 2018. "National Bureau of Statics of China Annual Data." http://www.stats.gov.cn/english/.

National Family Health Survey. 2016. *National Family Health Surveys Rounds 1, 2, 3 and 4*. http://rchiips.org/nfhs/.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge, UK: Cambridge University Press.

Panagariya, Arvind. 2008. *India: The Emerging Giant*. London: Oxford University Press.

Qian, Yingyi. 2017. *How Reform Worked in China: The Transition from Plan to Market*. Cambridge, MA: MIT Press.

Quah, Danny T. 1996. "Empirics for Economic Growth and Convergence." *European Economic Review* 40 (6): 1353–75.

Rajan, Raghuram G. 2012. "What Happened to India?" *Project Syndicate*, June 8. https://www.project-syndicate.org/commentary/what-happened-to-india?barrier=accesspaylog.

Rajan, Raghuram G. 2018. "Note to Parliamentary Estimates Committee on Bank NPAs." https://www.thehindubusinessline.com/money-and-banking/article24924543.ece/binary/Raghuram%20Rajan%20Parliamentary%20note%20on%20NPAs.

Ray, Debraj. 2010. "Uneven Growth: A Framework for Research in Development Economics." *Journal of Economic Perspectives* 24 (3): 45–60.

Redding, Stephen J., and Esteban Rossi-Hansberg. 2017. "Quantitative Spatial Economics." *Annual Review of Economics* 9 (1): 21–58.

Reserve Bank of India. 2019. *Handbook of Indian Statistics*. https://www.rbi.org.in/Scripts/publications.aspx.

Rodrik, Dani. 2016. "Premature Deindustrialization." *Journal of Economic Growth* 21 (1): 1–33.

Rodrik, Dani, and Arvind Subramanian. 2005. "From 'Hindu Growth' to Productivity Surge: The Mystery of the Indian Growth Transition." *IMF Staff Papers* 52 (2): 193–228.

Roy, Tirthankar. 2011. *The Economic History of India 1857–1947*. 3rd ed. London: Oxford University Press.

Sen, Amartya. 1990. "More Than 100 Million Women Are Missing." *New York Review of Books*, December 20. https://www.nybooks.com/articles/1990/12/20/more-than-100-million-women-are-missing/.

Shafik, Nemat, and Sushenjit Bandyopadhyay. 1992. *Economic Growth and Environmental Quality: Time-Series*

and Cross-Country Evidence. Washington, DC: World Bank.

Sitapati, Vinay. 2018. *The Man Who Remade India: A Biography of P. V. Narasimha Rao*. London: Oxford University Press.

Subramanian, Arvind. 2018. "India's Path from Crony Socialism to Stigmatized Capitalism." *Project Syndicate*, February 8. https://www.project-syndicate.org/commentary/india-cronyism-to-capitalism-by-arvind-subramanian-2018-02?barrier=accesspaylog.

Subramanian, Arvind. 2019a. "India's GDP Mis-estimation: Likelihood, Magnitudes, Mechanisms, and Implications." Center for International Development Faculty Working Paper 354.

Subramanian, Arvind. 2019b. "Validating India's GDP Growth Estimates." Center for International Development Faculty Working Paper 357.

Vanhanen, Tatu. 2012. "The Polyarchy Dataset, Version 2.0." https://www.prio.org/Data/Governance/Vanhanens-index-of-democracy/.

Varshney, Ashutosh. 1998. "India Defies the Odds: Why Democracy Survives." *Journal of Democracy* 9 (3): 36–50.

Varshney, Ashutosh. 2002. *Ethnic Conflict and Civic Life: Hindus and Muslims in India*. Cambridge, MA: Yale University Press.

Weber, Max. 1919. *Politics as a Vocation*. Berlin: Duncker & Humblot.

Weiner, Myron. 1991. *The Child and the State in India: Child Labor and Education Policy in Comparative Perspective*. Princeton: Princeton University Press.

World Bank. 2019. World Development Indicators. http://datatopics.worldbank.org/world-development-indicators/.

World Health Organization. 2019. "WHO Global Ambient Air Quality Database." https://www.who.int/airpollution/data/cities/en/.

Zaveri, Esha, Danielle S. Grogan, Karen Fisher-Vanden, Steve Frolking, Richard B. Lammers, Douglas H. Wrenn, Alexander Prusevich, and Robert E. Nicholas. 2016. "Invisible Water, Visible Impact: Groundwater Use and Indian Agriculture under Climate Change." *Environmental Research Letters* 11 (8): 084005.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 31–54

Why Does the Indian State Both Fail and Succeed?

Devesh Kapur

The most striking fact about the Indian state is how varied its performance has been, spanning the spectrum from woefully inadequate to surprisingly impressive.

On one side, Lant Pritchett (2009) memorably characterized India as a "flailing state." He wrote: "Measures of the administrative capacity of the [Indian] state on basics like attendance, performance, and corruption reveal a potentially 'flailing state' whose brilliantly formulated policies are disconnected from realities on the ground." India's state performs poorly in basic public services such as providing primary education, public health, water, sanitation, and environmental quality. While it is politically effective in managing one of the world's largest armed forces, it is less effective in managing public service bureaucracies. The research literature on India has many discussions of programs that fail to deliver meaningful outcomes, or that are victims of weak implementation and rent-seeking behavior of politicians and bureaucrats, or that are vitiated by discrimination against certain social groups (Niehaus and Sukhtankar 2013; Fisman, Schulz, and Vig 2014; Sheahan et al. 2018; Lehne, Shapiro, and Vanden Eynde 2018). For a comprehensive review of research on corruption in India, see Sukhtankar and Vaishnav (2015).

But on the other side, the Indian state has a strong record in successfully managing complex tasks and on a massive scale. It has repeatedly conducted elections for hundreds of millions of voters—nearly 900 million in the 2019 general elections—without national disputes. In this decade, it has scaled up large programs

■ *Devesh Kapur is the Starr Foundation South Asia Studies Professor, Paul H. Nitze School of Advanced International Studies (SAIS), Johns Hopkins University, Washington, DC. His email address is dkapur1@jhu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.31.

such as Aadhaar, the world's largest biometric ID program (which crossed one billion people enrolled within seven years of its launch). Most recently, it has implemented the integrated Goods and Services Tax (GST), one of the most ambitious tax reforms anywhere in recent times. India ranks low on its ability to enforce contracts, but its homicide rate has dropped markedly from 5.1 in 1990 to 3.2 (per 100,000) in 2016 (UNODC 2019).

The first section of this paper elaborates on the performance of the Indian state, highlighting the key paradox of its effectiveness on some difficult functions amidst its woeful inadequacies on others. It argues that the Indian state has delivered better in certain situations and settings: specifically, on macroeconomic rather than microeconomic outcomes; where delivery is episodic with inbuilt exit, rather than where delivery and accountability are quotidian and more reliant on state capacity at local levels; and on those goods and services where societal norms and values concerning hierarchy and status matter less, rather than in settings where these norms and values—such as caste and patriarchy—are resilient.

The second section proposes several explanations of these patterns of failure and success: understaffing of local governments, consequences of India's precocious democracy, and the persistence of social cleavages in India by caste, gender, and religion. A third section discusses two explanations for the poor performance of India's government that are often mentioned, but seem unlikely on further exploration: the claims that India's state sector is bloated in size and submerged in patronage. The conclusion offers some brief thoughts about some changing patterns of India's state capacity, which has seen notable improvements in its erstwhile weakness at the micro level even as its macro performance has become more worrisome.

Heterogeneous Performance

There is a vast literature on defining the role of the state and its effectiveness (for a recent overview, see Bardhan 2016). One strand has emphasized the role of the state in protecting property rights and in imposing constraints on itself so that the state does not become an instrument of expropriation. In the context of development, another strand has focused on the importance of an effective state for meeting development goals and reducing poverty (in this journal, Page and Pande 2018). Here, we focus more on the positive role of the state delivering essential services such as economic stability, health, education, regulation, and so on.[1]

[1] Rodrik and Subramanian (2003) provide a functional definition of the roles of the state in relation to markets: *market-creating*, which is providing rule of law and protection of property rights and ensuring sanctity of contract; *market-stabilizing and correcting*, which involves sound central banking and robust regulatory agencies; *market-legitimizing*, which involves tax and redistributive policies and affirmative action, providing voice and facilitating political participation; *market-complementing*, which is provision of public goods such as infrastructure and human capital; and *market-undermining*, through excessive interference in the form of state ownership on means of production and command and control policies.

Table 1
Changing Level and Structure of Tax-GDP Ratios in India

	Central taxes			*State's tax revenues*				
Year	*Direct*	*Indirect*	*Total*	*Direct*	*Indirect*	*Total*	*Total*	*Direct taxes/ total taxes (%)*
1950–1951	1.23	2.20	3.43	0.99	1.61	2.60	6.03	36.8
1965–1966	1.62	4.56	6.19	0.92	3.02	3.94	10.13	25.1
1989–1990	1.21	6.44	7.65	1.01	6.81	7.83	15.48	14.3
2016–2017	3.46	3.54	7.00	2.26	8.56	10.82	17.82	32.1

Source: Ministry of Finance (2018, table 1.8).
Note: "State's tax revenues" include the share of central taxes that are devolved to states.

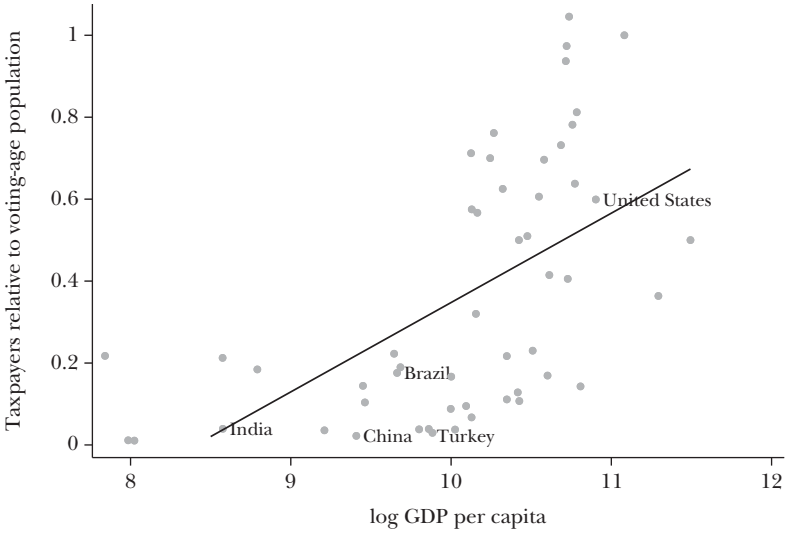
Fiscal Outcomes

At the heart of state-building is a fiscal story: just as a sovereign has monopoly over legitimate violence, it seeks to have a monopoly on legitimate expropriation of resources from the public in the form of taxes. This is especially the case for revenues from direct taxes (as distinct from indirect taxes) or revenues from natural resources or foreign aid (Brautigam, Fjeldstad, and Moore 2008; Besley and Persson 2013). Explanations of poor governance and limited service provisions are intrinsically linked to a state's limited ability to tax citizens: when few citizens pay taxes, they are less likely to feel a sense of ownership of the state, to demand services and accountability, or to punish corrupt practices (Persson and Rothstein 2015).

Pre-Independence British India was a strong state if measured by military capacity and a monopoly in the exercise of violence, but a weak state when measured by revenues (and concomitant expenditures on public goods). Government revenue as a proportion of national income was 2 percent in 1871 and only marginally higher at 3–5 percent in 1920–1930, compared to 19 percent in Britain and 29 percent in Japan in the interwar years (Roy 2011). On a per capita basis, between 1920 and 1930, British colonies in the Federated Malay States spent on average more than ten times the money spent in British India, that of Ceylon spent more than three times, other colonies such as the Philippines and the Dutch East Indies spent more than double, and French Indochina spent 40–50 percent more (Roy 1996).

India's weak fiscal inheritance improved after Independence in 1947, but only to a limited extent, with the tax-to-GDP ratio climbing from 6 percent in 1950–1951 to almost 18 percent in 2016–2017, as illustrated in Table 1. India's strategy of seeking to industrialize through import substitution from the 1950s into the 1980s led it to rely increasingly on indirect taxes (excise and trade taxes). As a result, direct taxes as a ratio of total taxes fell from 36 percent in 1950–1951 to just 14 percent in 1989–1990. Since then, the ratio climbed to 36 percent in 2007–2008 before declining to 32 percent in 2016–2017, lower than after Independence. Indeed, India's reliance on indirect taxes is likely to increase further now that the Goods and Services Tax has come into effect.

Figure 1A
Taxpayers Relative to Population



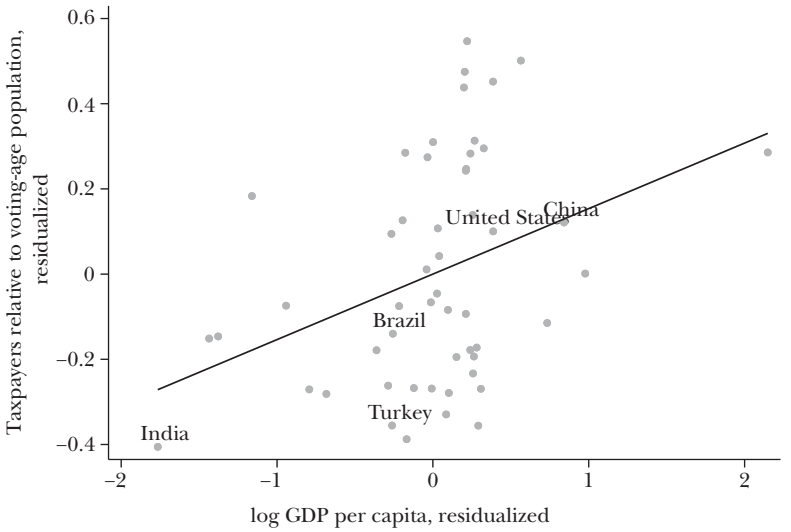
Source: Centre for Tax Policy and Administration (2011).
Note: The sample has 52 countries. The OECD dataset, which is the basis of this graph, has 77 countries. Of these, two countries do not have Polity IV data, and of the remainder, only 53 countries have data on the number of taxpayers and one country does not have data on GDP, leaving a final sample of 52.

Fiscal capacity generally improves with economic development. India's overall tax effort (measured as the number of income taxpayers) appears consistent with its level of development (as measured by per capita GDP), as shown in Figure 1A. [2] The results are similar if tax effort is measured simply as the tax/GDP ratio.

However, India is a democracy, and democracies typically tax and spend more than nondemocracies, likely because of the redistributive pressures they face. Acemoglu and Robinson (2000) argue that the extension of the franchise can be viewed as a credible precommitment to redistribution. Figure 1A shows the relationship between the ratio of the number of taxpayers to the population in a country and the log of per capita GDP. The solid line is simply the line of best fit. Figure 1B depicts the same relationship but controlling the level of democracy as measured by the average Polity IV score.[3] From this perspective, India is

[2]The discussion here draws upon India's *Economic Survey 2015–16* (Ministry of Finance 2016), where the tax-income-democracy relationship was first highlighted.
[3]To demonstrate this multivariate relationship on a figure, I appeal to the Frisch-Waugh theorem. I get the residuals from two regressions: (1) of the taxpayers to population ratio on Polity IV score and (2) of the log GDP per capita on Polity IV score. The figure plots the residuals from (1) on the residuals of (2). The Frisch-Waugh theorem guarantees that the slope of the regression line is the same as from a multivariate regression of taxpayers to population ratio on log GDP per capita and Polity IV score. The axes, therefore, plot the residualized values. I am grateful to Shoumitro Chatterjee for this analysis.

Figure 1B
Taxpayers/Population Controlling for Democracy Score



Source: Centre for Tax Policy and Administration (2011).
Note: The sample has 52 countries. The OECD dataset, which is the basis of this graph, has 77 countries. Of these, two countries do not have Polity IV data, and of the remainder, only 53 countries have data on the number of taxpayers and one country does not have data on GDP, leaving a final sample of 52.

a negative outlier. It has the lowest per capita GDP of countries that have been consistently democratic over the last half-century. In addition, India's tax effort—here measured by the number of taxpayers to per capita GDP—is substantially below what one might expect for democracies. (The result is similar if tax effort is measured by the tax/GDP ratio.) That is, India does not undercollect taxes because it is relatively low income; rather, it undercollects taxes, despite being a democracy. The patterns shown for taxes and taxpayers hold for expenditures as well.

India also displays striking differences in fiscal capacity across tiers of government. India is an outlier in that its subnational governments (many of which have larger populations than most countries) generate a very low share of their total revenues from direct taxes: about 6 percent compared to 19 percent in Brazil in 2016 and 44 percent in Germany (Ministry of Finance 2018). The reliance of India's rural local governments on its own resources is just 6 percent (compared to 40 percent for third-tier governments in Brazil and Germany), and they raise a meager 4 percent of their overall resources as direct taxes (compared with about 19 and 26 percent in Brazil and Germany, respectively). As a result, central and state governments in India spend on average 15–20 times more per capita than do local governments (Ministry of Finance 2018).

Macroeconomic Outcomes

India's record on delivering economic growth for the last four decades has been noteworthy in its duration, stability, and relatively high rate (as discussed in the article by Lamba and Subramanian in this symposium). The state adopted market-undermining policies in the first three decades after Independence in 1947, but corrected course beginning around 1980, and especially after 1991, to deliver an uncommon dynamism relative to its own past, as well as compared with other developing counties.

India's record on macroeconomic stability is evident in lower inflation, lower levels of external debt, and more conservative monetary policies than comparator countries. Its fiscal policy has, however, been less prudent: its average levels of fiscal deficits have been quite high even compared to other emerging market countries (according to IMF data), although overall debt levels are much more prudent. Joshi and Little (1994) show that prior to 1991, India's macroeconomic policies were the most "conservative" (amongst 17 countries studied) with respect to inflation, monetary policy, and external debt. India's inflation did ratchet up in the mid-2000s, reaching 10–11 percent by 2008, and remained elevated at double digits for several years. However, inflation in India has since then fallen sharply to 3–5 percent in 2018–2019. India's exchange rate management has also been prudent, moving from fixed to flexible exchange rates over time and avoiding bouts of overvaluation that have elsewhere led to currency crises. As per one summary measure of macroeconomic performance in developing countries, India has not been under an IMF program in the quarter-century since 1995. In addition, as of November 1, 2019, India's foreign exchange reserves exceeded $440 billion, the sixth highest of any country in the world.

India's fairly good inflation record owes much to the political aversion to inflation institutionalized by democracy. The very large number of poor people directly experience the negative impact of inflation, in part, because they lack access to financial instruments to protect themselves against inflation and also have high rates of electoral turnout. Thus, despite the absence of conventional statutory independence for India's central bank (the Reserve Bank of India), democratic politics has helped anchor and rein in inflationary expectations.

Microeconomic Outcomes Relating to Key Services

India has experienced considerable improvements in socioeconomic indicators in the seven-odd decades since Independence. Table 2 presents changes in some basic indicators. Table 3 focuses on the time period since the onset of India's economic liberalization in the early 1990s.

One potential question about these gains is whether some of the improvements mask poor quality. For example, a household might have access to electricity or drinking water, but this does not say anything about the quality of the service. This point has some validity, but on the other hand, there is little to indicate that quality was terrific earlier and quantitative expansion has largely come at the expense of quality.

Table 2
Some Socioeconomic Indicators in India

	~1950	*2016*
Literacy (% of population)	18[a]	76[d]
Life expectancy (years)	32[a]	68.7[e]
Maternal mortality (per 100,000)	1,321[b]	130[e]
Real GDP per capita (in 2011 US dollars)[c]	824	6,125

Source: [a]India Census Commissioner (1951); [b]estimates are for 1957–60 (Radkar 2012, 120, table 1); [c]Bolt et al. (2018); [d]extrapolated from National Sample Survey Office (2015), which estimated literacy for age five and above at 76 percent; [e]CBHI (2019).

Table 3
Post-liberalization Changes in Social Outcomes

Indicator	1998–1999 (NFHS-2)	2005–2006 (NFHS-3)	2015–2016 (NFHS-4)
Households with electricity (%)	60.1	67.9	88.2
Households with an improved drinking-water source (%)	77.9[a]	87.9	89.9
Infant mortality (deaths/1,000 live births)	68	57	41
Female literacy (%)	41.8	55.1	68.4
Women with ten or more years of schooling (%)	14.3	22.3	35.7
Sex ratio at birth (number of females per 1,000 males for children born in the last five years)	926[b]	914	919
Homicide rate (per 100,000)[c]	4.6	3.9	3.2

Source: National Family Health Surveys (various rounds); [a]refers to drinking water supply from piped water and hand pumps; [b]0–6 years; [c]UNODC (2019).

A pattern that emerges here is that India's progress is generally better in areas where the state can deliver the service by planning, coordination, and financing, such as electrification or road connectivity. In 2015, 88 percent of India's population had access to electricity, a substantially higher percentage than the 66 percent that would be expected based on a simple cross-country correlation between electrification and GDP per capita. However, where behavioral changes on "sticky" social norms and preferences are required, India's progress has been slower. For example, 39 percent of India's population practiced open defecation in 2015, also a substantially higher percentage than the 14 percent that would be expected based on a simple correlation between this practice and GDP per capita.[4] When it comes to issues related to women and children's welfare, some of them shown in Table 3, the Indian state has been less effective. India's adverse sex ratio reflects society's

[4]For figures illustrating these cross-country correlations with access to electricity and open defecation, see the online Appendix available with this paper at the *Journal of Economic Perspectives* website.

strong son preferences, and despite legal proscriptions, there have been meager improvements.

Episodic versus Ongoing Delivery of Key Functions

The Indian state performs better in activities that are episodic in delivery and accountability and where, therefore, exit is automatic once the activity is complete. Consider three constitutional and statutory bodies charged with crucial functions. India's elections are organized by the Election Commission of India, which normally has a bureaucracy of a few hundred people. During national elections, it has supervisory authority over all parts of the bureaucracy be it national, state, or local—several million officials, a ten-thousand-fold increase—but only for the period of time between when the election schedule is announced and election results declared (usually a couple of months) (Quraishi 2019). Another body, the Delimitation Commission, is set up periodically to reallocate parliamentary and state assembly constituencies based on the last census and is wound up once it submits its report to parliament.[5] Each of its reports has been adopted unanimously by an otherwise fractious parliament, allowing India to avoid the gerrymandering partisanship that has afflicted, for instance, many US states. India's fiscal federalism—the sharing of taxes both vertically and horizontally—is strongly guided by the Finance Commission, which is set up every five years and then wound up after making recommendations, which again have been largely accepted by the national and state governments of the day.

The better performance of the Indian state in time-bound activities with automatic exit is not just with regard to statutory bodies. Earlier in 2019, the Indian state's organization of the Kumbha Mela—the "world's largest human gathering" now recognized by UNESCO as "an intangible cultural heritage of humanity"—involved the construction of a temporary city (at Prayagraj) spread across 2,500 hectares. Around 220 million people attended the 50-day festival with more than 10 million participating on the final day, all without any serious mishaps.[6]

Public health services in India leave much to be desired. Yet India achieved a remarkable public health milestone when it completed a full five years as a "polio-free nation" on January 13, 2016. Even into the 1980s, tens of thousands of children were contacting polio each year. As late as 2009, India reported 741 polio cases, more than any other country in the world. It faced daunting challenges in eradicating polio: high population density and birth rate, poor sanitation, widespread diarrhea, inaccessible terrain, and the reluctance of a section of the population to accept the polio vaccine. The sheer scale of the effort, requiring 172 million children to be vaccinated twice each year, all within a day or two, with the assistance of about 2.5 million volunteers and 150,000 vaccine administration

[5] Under Article 82 of the Constitution, the Parliament by law enacts a Delimitation Act after every census.
[6] For comparison, the much richer Saudi state organizes the Hajj religious pilgrimage for about 2.4 million people spread over five to six days.

supervisors, required substantial state capacity in logistics and coordination.[7] Again, the Indian state performed well in a "mission mode" activity that was highly temporally concentrated.

The Indian census is another example. It has been held on time every decade. When the 2011 census was conducted, 2.7 million officials visited households in 7,935 towns and 600,000 villages at a cost of less than $0.50 per person. As with the earlier examples, the activity is conducted by the same public employees who don't seem to perform well on their otherwise normal day-to-day duties (indeed, the census results take quite a while to be processed), but do better on an episodic time-bound activity with automatic exit as soon as the activity ends.

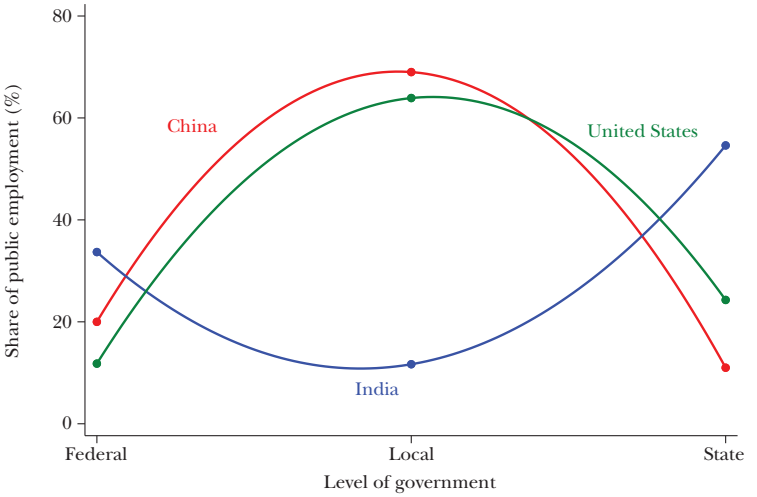
When it comes to ongoing regulatory agencies, however, the performance of India's state is more mixed. Primary financial regulators like the Reserve Bank of India and the Securities and Exchange Board of India were generally regarded as having performed well, at least until recently. However, India's current deep financial crisis rooted in high nonperforming assets of Indian banks raises serious questions about the former. The record of the electricity regulator is more modest, environmental regulation has not been successful, and the telecommunications regulator began well but has now come under a cloud for seemingly favoring a particular firm (Kapur and Khosla 2019). Regulation at the central level is also generally better than by counterparts at the subnational level.

Since the onset of economic liberalization, the role of the Indian state in the country's economy has shifted from direct production to regulation. However, the resulting profusion of regulators has led to increased fragmentation of the state apparatus and further slowed decision-making. For example, much needed reforms in agricultural marketing have been stymied because implementation has to be done by each state separately, since agriculture is a "state subject" in the Indian Constitution. Indeed, if "economic liberalization was meant to roll back the state, the rise of the regulatory state is a testament to the fact that, far from rolling back, the state has simply rolled over" (Kapur and Khosla 2019, 5).

This regulatory turn has furthered the "judicialization" of the Indian state, since all disputes end up in that branch (Mehta 2007). India's judiciary has emerged as a critical feature of state capacity; this is in addition to its traditional role of being a referee institution that protects rights and checks the other branches of government. In response to the weakened inability and legitimacy of the executive and legislature and other regulatory bodies to take actions, the judiciary has stepped into the vacuum by appropriating roles that typically are not within its jurisdiction. Some examples of economic issues that are decided in part by the judiciary include allocation of public assets, standard-setting for environmental quality, regulating utilities, and driving the bankruptcy process. Even where decisions have to be made by other regulatory institutions, they are

[7] See Polio Global Eradication Initiative (2016).

Figure 2
Structure of Employment across Three Tiers of Government: China, India, and the United States



Source: India: Ministry of Finance (2012, 2018); China: Ang (2012); United States: US Census Bureau.
Note: The data are for 2011–2012 for India, 1998 for China, and 2012–2013 for the United States.

routinely "punted up" to India's Supreme Court, which until recently at least has been consistently ranked as one of the two most trusted public institutions (along with the army) (CSDS 2015).

Possible Explanations

What explains India's pattern of heterogenous state delivery? In this section, we consider several possible explanations; in the following section, we rule out two conventional but implausible explanations.

Explanation 1: Inadequate Local Government Resources

There is a striking contrast in the structure of public employment in India at different levels of government compared to other large federal states, like the United States and China, which is illustrated in Figure 2.

In the United States, two-thirds of government employees work for local governments. In the case of China, from 1954 to 1998, "central-level public employment averaged 16 percent of the total, provincial level 17 percent, city level 22 percent, county level 34 percent and townships 10 percent. On average, 66 percent of public employment [was] at the subprovincial levels, with county governments having the largest share" (Ang 2012, 693). From 1980–1998, central-level staff in China grew from 12 percent to 20 percent, while the share of provincial governments declined

from 18 percent in 1979 to 11 percent in 1998, reflecting in part a steady devolution of responsibilities from provincial to subprovincial administrations (Ang 2012).

In contrast, in India during the period from 1980 to 2012, the share of the central government declined from 21.1 to 14.3 percent, that of local governments from 13.8 to 12.0 percent, while the share of subnational government increased from 36.3 to 40.8 percent.[8] Thus, the share of local government employees in total employment in the United States or China is five times that of India.

Local government expenditure is similarly skewed. Local government expenditure is 3 percent of the total government expenditure in India, compared with 27 percent in the United States and 51 percent in China (Ren 2015). Given that basic public goods—from primary health to education, from water to sanitation, from policing to (urban) planning—are supplied (or should be supplied) by local governments, poor delivery of many basic services in India could relate to the lack of resources (both financial and human) at the lowest level of government.

We lack a good understanding of how the vertical distribution of public employees across levels of government in large federal states might shape state effectiveness, or for that matter, how the horizontal distribution of personnel across ministries in government affects state performance. No matter how carefully development programs are designed by national bureaucracies, ultimately their performance on the ground hinges on how effectively they are implemented by local bureaucracy at the front line (Pritchett 2009). Bureaucratic resource constraints affect performance by forcing local officials to multitask excessively, and this inability to specialize has an adverse impact on the performance of development programs (Dasgupta and Kapur 2019).

Explanation 2: Precocious Democracy

India's democratic persistence has defied theorizing about democracy. One well-regarded study on the relationship between democracy and development found that India was a major outlier given its low level of income and literacy and high levels of ethnic and religious conflict: "India was predicted a dictatorship during the entire period … the odds against democracy in India were extremely high" (Przeworski et al. 2000). In particular, when India's constitution guaranteed universal franchise for men and women in 1950, real GDP per capita in India was lower than what Western democracies like the United Kingdom, United States, Sweden, or the Netherlands had a century or more before (when of course universal franchise was unknown). India adopted the universal franchise when literacy was barely 18 percent and life expectancy was just 32 years.

In addition, national-level democracy and the universal franchise in India arrived all at once in 1950. The expansion of the franchise in most nations of the

[8]The balance of public employment is under "quasi-government," which includes state-owned enterprises, sundry authorities, boards, regulators, etc. The data are from the annual Economic Surveys of the Government of India for the years 2017–2018, 2011–2012, 2001–2002, 1995–1996, 1984–1985, 1974–1975, and 1972–1973.

West was much more gradual, from male property holders to all men to women and later (in some cases) to members of marginalized communities. In countries of East Asia, there was a clear sequencing with universal franchise following a protracted process of economic development and state-building. Most Latin American countries experienced periodic reversals in democracy even as their economies grew. In both Western democracies and East Asia, universal franchise came after the state had laid the foundations of public goods in education and health, and the structures of the welfare state were built gradually on these foundations.

Might the weaknesses of India's state record be the outcome of a precocious democracy, arriving much sooner than one might expect based on comparative experience? We suggest that the distinctiveness of Indian democracy is responsible for the heterogeneity of its government performance in three distinct ways.

First, precocious democracy tends to militate against the provision of public goods in favor of redistribution. Countries that experienced economic development prior to the transition to democracy also tend to adopt democratic institutions that constrain the confiscatory power of the ruling elite. However, when countries pursue democracy prior to economic development, the democratic institutions adopted enhance the redistributive powers of the state.

For related reasons, precocious democracy contributes to weak public good provision. We suggested earlier that being a "premature" democracy appears to have reduced India's ability to raise revenues compared with other democracies, further undermining its fiscal ability to finance public goods. By not providing public goods before shifting to redistribution, the Indian state weakened the legitimacy and trust to create a virtuous circle that could strengthen the social contract between citizens and the state. This has led to "exit" (in the sense of Hirschman 1970): India's middle class, feeling that it has not received enough from the state by way of public goods, exits in favor of private provision and is also reluctant to pay taxes. This further undermines the state's legitimacy and capacity (reflected in the low level of taxpayers in India). In an Indian twist to these pathologies, exit and lack of trust and weak fiscal ability are particularly acute at lower levels of government, which have the primary responsibility for delivering key services such as health, education, water, and sanitation. There is a general unwillingness on the part of lower levels of government to raise revenues and local taxes (such as those on property)—a problem of being closer to the people. This creates a vicious cycle of weak delivery leading to exit, undermined legitimacy, fewer resources, back to weak delivery.

Redistributive pressures can also explain variance in regulatory effectiveness. The electricity sector is a clear example where politicians press to keep charges for farmers and households low and attempt to cross-subsidize by charging higher for commercial and industrial units, creating severe distortions in the sector.

Second, a precocious democracy with electoral mobilization along social cleavages favors creation of narrow club goods. A central puzzle concerning the poor provision of basic public services in India is seemingly weak demand in an otherwise flourishing electoral democracy. If politicians respond to voters, then why have

voters not demanded basic public services such as education, health, water, and sanitation? On this point, explanations seek recourse to the implications of India's social heterogeneity, especially in a ranked society (Banerjee, Iyer, and Somanathan 2005). Politicians persist in relying on targeted transfer programs and subsidies. Poor farmers might prefer targeted transfers rather than public services such as education because farmers often have high discount rates (Keefer and Khemani 2004). Alternatively, politicians might provide "private" public goods such as housing or other material inducements that target particular individuals and small groups of people as opposed to the provision of public goods through institutional means. This behavior is partly due to social cleavages and partly to a lack of credibility of political promises to provide broad public goods (as opposed to private transfers and subsidies). Electoral competition therefore revolves around distributing public resources as "club goods"—goods with excludability characteristics—rather than providing public goods to a broad base. Again, those in the middle and upper classes will often choose to exit from the system, preferring market solutions rather than poor quality and unreliable public services, further reducing pressures to change the system.

A third way in which precocious democracy can weaken state capacity is that an imperfect democracy with noncredible politicians will tend to emphasize the provision of goods that are visible and can be provided quickly, like infrastructure, over long-term investments, like human capital or environmental quality. This pattern relates to the signaling problem facing noncredible politicians in an electoral democracy and the politics of visibility (Mani and Mukand 2007). There is a bias towards tackling famines over addressing malnutrition (Dreze and Sen 1989) because the former are more visible. There is also a bias towards public goods where quantity is more salient than quality (Hirschman 1967). While this pattern may hold in many democracies, Indian democracy might be particularly susceptible because of lower levels of literacy and social cleavages where the politics of visibility and "signaling" to specific groups becomes more salient.

Explanation 3: The Persistence of Social Hierarchies and Cleavages

The architect of the Indian Constitution, B. R. Ambedkar, was born into an "untouchable" caste. Ambedkar worried that "democracy in India is only a top-dressing on an Indian soil which is essentially undemocratic."[9] Caste, he argued, "is a notion, it is a state of the mind," and therefore it cannot be eradicated through constitutional measures alone. This "state of mind" has been embedded in India's tenacious social institutions. These constitute what one could call "societal failures" and make state failure more likely in areas where public policy and programs have to address issues intimately connected with caste or gender issues.

The framers of India's constitution sought to construct an institutional form "that could counteract the tenacity of local cultural forms," and for this reason, they

[9] See Constituent Assembly Debates (1948).

sought to concentrate authority at the center (Khosla, forthcoming). To distance power from local actors meant (they believed) that power could be exercised progressively, since the boundary between the state and society became porous as political authority traveled downward. With hindsight, we now know that many of these egalitarian hopes were misplaced. Policies—often largely progressive in intent—were made by national elites who were more insulated from society. Implementation, however, was often subverted by local elites in what was an extremely hierarchical society—and over seven decades has gradually become only somewhat less so.

While the Indian Constitution banned caste discrimination, it continues to be a social and political reality. Historically, the critical mechanism for replicating caste distinctions across generations is endogamy, with people marrying within their caste. Ambedkar (1936) had argued, "Where society is already well-knit by other ties, marriage is an ordinary incident of life. But where society is cut asunder, marriage as a binding force becomes a matter of urgent necessity. The real remedy for breaking Caste is intermarriage. Nothing else will serve as the solvent of Caste." But to what degree can a democratic state interfere in marriage choices?

Societal failures are also manifest in the reality of gender discrimination in India. The World Economic Forum's 2018 Gender Gap Index places India at 108 out of 149 countries. Its rank on health and survival was 147 and in educational attainment 142. Its female labor force participation rate fell from 42.7 percent in 2004–2005 to 23.3 percent in 2017–2018, one of the lowest in the world outside the Arab world.

To take another example, a recent flagship program (the Pradhan Mantri Ujjwala Yojana) has provided gas connections to poor households to facilitate women moving away from highly polluting and time-consuming solid fuels. Launched in May 2016, the scheme had provided cooking gas connections to 80 million poor households by late 2019. But many women continue to cook using solid fuels, even when a gas stove is available, despite the adverse health effects of solid-fuel use. While the cost of refilling a gas cylinder is an important consideration, beliefs and attitudes seem to matter, too. A recent survey found that over 85 percent of respondents saw solid fuel as a better option for taste and the health of family members eating the food (Gupta et al. 2019). In Indian households, the health of young women, who do most of the cooking, is less valued than that of other household members, who do most of the eating. A democratic state can ensure that households have healthier cooking fuel options, but it can only do so much to address what happens *within* the household—at least in the short term.

The scholarship on India that takes societal factors into account has focused more on how social heterogeneity can affect electoral incentives in a way which makes collective action more difficult (Bardhan 1986) and reduces incentives to invest in state capacity (Besley and Persson 2011). Several empirical studies have documented that India's social diversity has hindered development outcomes and, in particular, the provision of public goods (Banerjee, Iyer, and Somanathan 2005; Balasubramaniam, Chatterjee, and Mustard 2014).

If universal franchise results in the state representing the interests of the median voter (for example, as in Alesina and Rodrik 1994), the very success of India's electoral democracy means that the Indian state is also likely to reflect societal preferences on issues such as caste and gender. This explains why the Indian state is relatively better at providing "hard" goods such as roads or electrification rather than providing sanitation because dealing with human waste was seen as "polluting" and restricted to the lowest castes. Similarly, the state is better at raising food grain production (four-fold over the past half-century) than improving malnutrition outcomes, which are affected by intrahousehold distribution. It is better at building schools and giving bicycles to improve girls' enrollment than at improving educational outcomes, because what happens within the classroom is affected by caste and gender norms. India's state is even less effective in improving worrisome sex ratios, low (and declining) female labor force participation, or generalized societal violence against women—all of which are rooted to varying degrees in social norms. In these cases, state failures reflect societal failures.

Two Conventional but Unpersuasive Explanations

Unpersuasive Explanation 1: Bloated Size of the Indian State

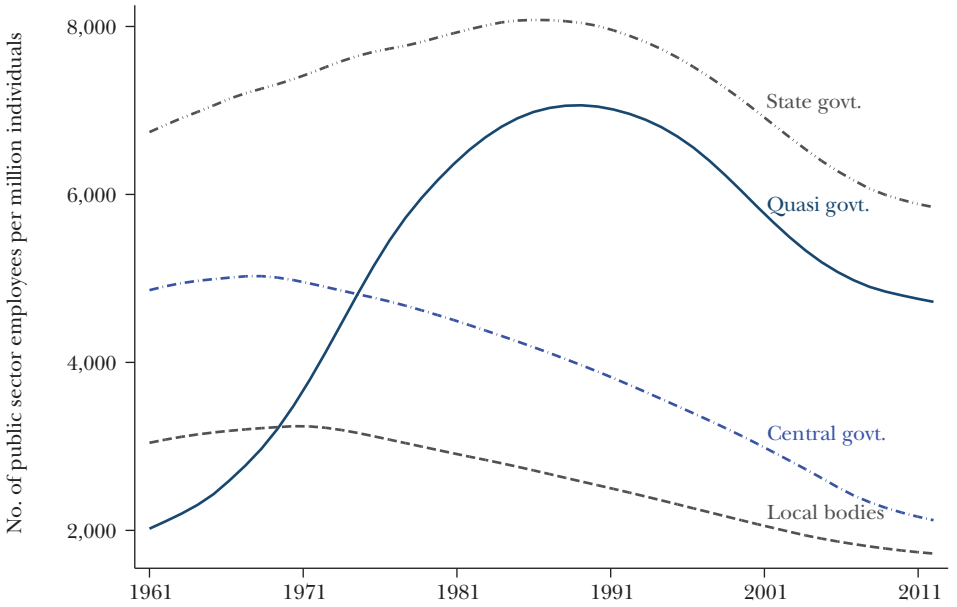
A conventional caricature about the Indian state is that its inefficiencies arise because it is too big, with overstaffed and lethargic bureaucracies doing little and seeking to extract rents. But if anything, the Indian state is relatively small in comparative context when measured by the number of personnel.[10]

In the early 1990s, the global average of government employment as a percent of population was 4.7 percent. In countries of Asia, it was 2.6 percent. In India, it was 2 percent (Schiavo-Campo, De Tommaso, and Mukherjee 1997). Core elements of the Indian state—police, judges, and tax bureaucracy—are among the smallest of the G-20 countries. Indeed, while the absolute size of government employment peaked in the mid-1990s, in relative terms, the decline in size of central and local governments began much earlier, as shown in Figure 3.

These numbers in fact overstate the size of India's federal government. A noteworthy aspect of the deployment of personnel in India's central government is that of its 3.24 million personnel (excluding military) in 2014, the Indian Railways and Posts alone employed 1.32 million. The Indian Railways constitute a very large fraction of central government employment; it accounted for 57 percent of total federal government employment in 1957, while its share was still 40 percent in 2014. In addition, nonmilitary security personnel (in the Ministry of Home Affairs and civilians

[10] Of course, it is possible that even though the number of public employees might be fewer, their emoluments could still be relatively large, resulting in a large share of government revenues being cannibalized by the state to feed its own employees. While salaries of public employees have been increasing at a faster pace relative to per capita income, expenses for compensation of public employees as the share of total government expenditures for India are relatively low in comparative terms.

Figure 3
Public Employment in India by Different Levels of Government
(per million population)



Source: Ministry of Finance (1973, 1975, 1985, 1996, 2012, 2018).

in the Defence Ministry) account for another 1.38 million personnel (as of January 1, 2014). Indeed, the most significant increase in the number of workers at the federal government level has been in the Ministry of Home Affairs from 325,000 in 1984 to 972,000 in 2014, due to the large expansion of central paramilitary forces.

At most, only 13 percent of federal employees in India—whose overall numbers are not large to begin with—are employed in core development-related departments. The number of personnel working in the secretariats of all ministries/departments of the central government was less than 30,000 in 2014.

A comparison between the size of the civilian workforce of the federal government in India with that of the United States is instructive. Remember, India has a large number of public enterprises and public sector banks that are under the central government. Even so, the size of the Indian federal government is half the size of its US counterpart when normalized by population: specifically, the US federal government had 8.07 civilian employees per 1,000 US population in 2014, down from 10.4 in 1995, while India's central government had 4.51 civilian employees per 1,000 population, down from 8.47 in 1995.[11]

[11] These estimates are based on various sources: US Office of Personnel Management (1940–2014), US Postal Service (2014), Government of India (2015), Ministry of Heavy Industries and Public Enterprises (2018), Resereve Bank of India (2018), and World Bank National Accounts data.

A second feature of the Indian state is the steady decline in public employment despite rapid growth. Higher income countries tend to have larger government expenditures as a share of GDP, and countries with higher per capita incomes tend to have larger state sector employment, especially in public services such as law and order and social services such as health care. However, in India, the total number of public employees at all levels of government rose from about 16,000 per million population in the early 1960s to a peak of around 19,000 per million population in 1986. However, since then, public employment has fallen at all levels of government, dropping to about 14,000 per million population by 2012, despite steady and strong increases in per capita income.

Unpersuasive Explanation 2: Patronage State

Another conventional explanation for poor performance is that the Indian state is a "patronage state," at least as regards recruitment into the public sector. However, recruitment to government jobs in India has become strongly rule based, largely (and increasingly) through exams. More than four-fifths of those who join India's federal civil service today are recruited through exams (and the rest through public advertisements and interviews conducted by an autonomous constitutional body, the Union Public Service Commission) at the federal level and equivalent bodies at the state level. When a new government takes power in the United States or Mexico, there is a substantial hiatus as new political appointees are selected for hundreds of senior-level positions. In India's case, there is a shuffling within the bureaucracy, but little lateral entry with very few new outside appointments.[12]

From the 1960s through the 1990s, about three-fifths of India's senior federal bureaucracy were recruited through exams. More than 2.9 million people took the federal civil service recruitment exam in 2015–2016, and 5,659 were recommended for appointment—roughly one of every 500. Of those taking the exam, the share recommended for appointment has fallen over time: back in 1950–1951, one out of ten test-takers was recommended.[13]

Recruitment at the subnational and local levels is more prone to bribes and patronage. But even here, recruitment in state public services is almost entirely exam based, conducted (depending on the cadre and the state) by State Public Service Commissions, Staff Selection Commissions, Professional Examination Boards, and the Uniformed Services Recruitment Board (for the police and fire services). There

[12] In order to bolster the generalist cadres of the civil services, the federal government recruited "domain experts" (those with demonstrated sectoral experience) in key policy positions (Joint Secretary) for the first time in April 2019 after much contentious debate. The recruitment was done by India's Union Public Service Commission, and the total number recruited was just nine.

[13] India's government is selective in recruitment in other ways as well. For example, the share of applicants accepted into the key training school for military officers in India (the National Defense Academy) is more than 200 times smaller than its US counterpart (West Point). The data for India are averaged over 2014–2016 and for the United States over 2017–2018. The sources for this comparison are the UPSC Annual Reports, 2014–2016, for India and the West Point Admissions Class Profile, Princeton Review for the United States.

are undoubted shenanigans (such as leaking of exam papers). Selections are often halted and sometimes overturned by the courts, and as a result, instead of overstaffing many positions are unfilled.[14]

The hypothesis of patronage recruitment as the principal explanation for India's poor state performance is also undermined by the large numbers of vacancies across all branches of the Indian state. About one-fifth of all positions in the central government are vacant; as well, one-third of all High Court judges and one-fifth of Supreme Court judges have been vacant at any given time this decade. About one-third of faculty positions in the elite Indian Institutes of Technology (which come under the federal government) are vacant. At the state level, about one-quarter of positions of district judges and police are vacant.

There are many possible reasons for the high number of vacancies in government jobs: lack of funding; certain judicial interventions might freeze government hiring; or for certain jobs, there are just not enough suitable candidates because of limited supply (as in the case of faculty vacancies for the elite Indian Institutes of Technology). High vacancies in certain areas could also be the result of more-or-less deliberate (in)actions by politicians. For instance, high vacancies among the police and judiciary ensures that law and order is weak. This not only ensures that politicians' shenanigans are unchecked, but that citizens have to seek out politicians to resolve disputes rather than the machinery of the state, making them more beholden to politicians.

However, for the most part, such factors cannot explain high vacancies across the gamut of the state machinery, such as the armed forces or agriculture officers. Between 2006 to 2014, average annual recruitment to India's central government was just above 100,000 annually, while the labor force was increasing by about 9 million each year.[15] These vacancies have persisted despite the obvious political imperatives to stack the state with partisan supporters, along with acute joblessness as a salient political issue. For the most part, the potential supply of government workers seems overwhelming relative to demand.[16] In a study of the Office of the Block Development Office (BDO), the key local-level administrative office serving about a quarter of a million people in rural India, Dasgupta and Kapur (2019) find that on average 48 percent of officially sanctioned full-time employee posts were vacant. If all of these sanctioned but vacant positions were filled, they find that the performance of one of the flagship government programs, the National Rural Employment Guarantee Act (NREGA), would increase employment delivery by approximately 10 percent.

[14] It is possible that those opting for public service may have a higher propensity for corruption, which may undermine the positive selectivity on ability (Hanna and Wang 2017).

[15] The data are from Annex 3, p. 45 in the *Report of the Seventh Central Pay Commission* (GOI 2015).

[16] In early 2019, a nationwide call for recruitment for menial positions in the Indian railways—porters, welders, and track maintainers—attracted 19 million applications for 63,000 posts. In 2018, 93,000 applicants vied for 62 peon jobs in the Uttar Pradesh police. Nearly 200,000 competed for 1,137 constable positions in the Mumbai police.

There are of course well-known examples of patronage in India. Absenteeism in certain types of jobs such as teachers and medical workers (Muralidharan et al. 2017) and an increase in "contract workers" (such as school teachers) is linked to rent seeking and political alignments, reducing state effectiveness. Another weakness that allows for the "patronage state" to flourish is the prevalence of a "transfer Raj," where officials in rent-rich posts are transferred at will depending on their ability to pay upwards in the food chain. Numerous studies have shown the negative consequences of transfers on the effectiveness of the bureaucracy in India ranging from irrigation (Wade 1982), teachers (Ramachandran et al. 2018), health workers (Purohit, Martineau, and Sheikh 2016), the elite civil service (Iyer and Mani 2012), and police (Das and Sabharwal 2017). There are cases where senior bureaucrats are offered high-profile post-retirement positions as regulators, and the performance of regulators appointed in this way has a checkered record.

But overall, the reality is that a significant part of the Indian state is served by a closed well-paid professional bureaucracy, recruited meritocratically through highly competitive formal examinations, with career stability and secure tenure, strong ties among the members of the bureaucracy, special laws for public employment (as opposed to standard labor laws), and internal promotion. These criteria would suggest that the core bureaucracy in India meets most of the conditions of a Weberian bureaucracy, in contrast with patronage states where political and personal criteria largely determine bureaucratic recruitment and careers (Rauch and Evans 2000; Dahlström, Lapuente, and Teorell 2012). Yet its performance has been wanting.

Some Concluding Thoughts

Between 1994 and 2019, India added 418 million people to its population. It added more people during these 25 years than its total population at Independence (361 million in the 1951 census), more people than the entire population of the United States—in one-third the US land mass. Yet the weak state in India, using democratic means, managed to enact a set of economic reforms in the early 1990s and to conduct economic policy since then in ways that laid the groundwork for sustained and robust economic growth in the last quarter-century.

The literature on the role of the state often yearns for a "Goldilocks" state: not too weak to be unable to formulate and implement policy in the larger public interest, but not too strong lest its "grabbing hand" undermine private property rights, markets, and contracts. This requires building a competent bureaucracy together with external checks and balances, including constitutional constraints on executive power, separation of powers, electoral rules, independent judiciary, free media, and various other factors.

However, as this paper has argued, India ostensibly has many of these institutional features, and yet state performance has left much to be desired. The Indian state has performed better in activities which are episodic, where the good or

service is a narrow club good and where a small technocratic bureaucracy suffices. It does less well where rents and social cleavages overlap. It does least well on issues that require behavioral changes at the micro level. The reasons, we argue, lie in the understaffing of local government, the precocious democracy of India and its anomalous sequencing of universal franchise, and India's "societal failures" manifest in caste and gender discrimination.

India's experience highlights a shortcoming in studies of state capacity; an emphasis on the "institution" aspect often ignores the "organizational" aspects of the state—and even that literature has for the most part focused on a narrow set of civil service management structures focusing primarily on incentive and monitoring (for an overview, see Finan, Olken, and Pande 2017). But there is often only limited analysis regarding how state performance is affected by public personnel management practices, such as the role of intrinsic motivation; pay structures (rather than levels); promotions; transfers; the composition of teams and physical conditions within which local bureaucracies work; and the distribution of bureaucracy across different functional lines as well as across different tiers of local, state, and federal government.

Perhaps the most intriguing trend with regard to the Indian state and the evolution of state capacity is an improvement at the level of micro and frontline implementation. However, there are growing questions about state capacity at the macro policymaking level.

The frontline implementation capacity of the Indian state is improving markedly and is manifest in its ability to scale up programs rapidly to reach tens and even hundreds of millions of people. In this decade, India's state has successfully opened bank accounts for over 350 million people, delivered gas connections to more than 80 million households, built around 100 million toilets reaching 600 million people, and has begun implementing direct cash transfer schemes that are reaching tens of millions of farmers. While each of these programs has exaggerated numbers and challenges of quality, timeliness, and exclusion, there is little doubt that the Indian state is now developing the capacity to transform inputs into outputs. Of course, transforming these outputs into outcomes is yet another step. Building toilets is not the same as usage, let alone sanitation writ large and better health outcomes. But it's a start.

Some of the "front-end" weaknesses of the Indian state are being attenuated by harnessing technology to implement programs on scale. The creation of a finance-biometric-communications platform Aadhaar, encompassing the entire population, and the development of Unified Payments Interface as the platform for digital payments are examples in this regard. Together with sharp improvements in connectivity, from rural roads to electrification and digital access, the platforms for improving the delivery of public programs, as well as for markets to function better, are strengthening.

But while technology has sharply reduced the transaction costs of obtaining a host of government documents, such as a passport or driver's license or to pay utility bills, it does not by itself get water into a house or sewerage out of it or treat it before discharge. For all of these, India will need a more effective state, one that

is better resourced especially at the local level and whose accountability is more "downward" directed towards citizens, rather than "upward" directed towards the state-level bureaucracy and politicians.

In contrast to the improvements at the micro level, the macro policy capabilities of the Indian state are raising concerns. Since India's Independence, observers have admired the state's capabilities embedded in its elite bureaucracy and its ability to formulate policy, while lamenting the severe weaknesses of its front-line functionaries to implement programs. A strong head and weak body, as it were, resulted in better macro performance and contrasted with the Indian state's mediocre implementation record on the ground.

Today, these patterns seem to be reversing. Even as the delivery of India's public programs has been improving, economic growth is stalling, if not declining (as discussed in the article by Lamba and Subramanian in this symposium). The autonomy of the core institutions of the Indian state and democracy—from the Supreme Court to the Election Commission, from the Reserve Bank of India to its statistical institutions—appears to be under growing political pressures. The lament about the Indian state as "weak" is becoming less true, but the same strong state that can ensure more effective poverty programs can also limit civil liberties and be less permissive for democracy. Sometimes one should be careful what one asks for—one might actually get it.

■ *I am grateful to Arvind Subramanian for his many stimulating ideas and insights that helped develop this paper and Shoumitro Chatterjee for innumerable discussions and technical help. I thank Pratap Bhanu Mehta, Sharun Mukand, Gulzar Natrajan, Lant Pritchett, Neelanjan Sircar, TV Somanathan, and Milan Vaishnav for helpful comments. Sumitra Badrinathan, Jashan Bajwa, Saksham Khosla, and Vaishnavi Rupavatharam provided excellent research assistance. I am grateful to the editors of this journal for extremely useful comments.*

References

Acemoglu, Daron, and James A. Robinson. 2000. "Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective." *Quarterly Journal of Economics* 115 (4): 1167–99.

Alesina, Alberto, and Dani Rodrik. 1994. "Distributive Politics and Economic Growth." *Quarterly Journal of Economics* 109 (2): 465–90.

Ambedkar, B. R. 1936. *Annihilation of Caste.* Lahore: printed by the author.

Ang, Yuen Yuen. 2012. "Counting Cadres: A Comparative View of the Size of China's Public Employment." *China Quarterly* 211: 676–96.

Balasubramaniam, Divya, Santanu Chatterjee, and David B. Mustard. 2014. "Got Water? Social Divisions

and Access to Public Goods in Rural India." *Economica* 81 (321): 140–60.

Banerjee, Abhijit, Lakshmi Iyer, and Rohini Somanathan. 2005. "History, Social Divisions, and Public Goods in Rural India." *Journal of the European Economic Association* 3 (2–3): 639–47.

Bardhan, Pranab. 1986. *The Political Economy of Development in India*. Delhi: Oxford University Press.

Bardhan, Pranab. 2016. "State and Development: The Need for a Reappraisal of the Current Literature." *Journal of Economic Literature* 54 (3): 862–92.

Besley, Timothy, and Torsten Persson. 2011. *Pillars of Prosperity: The Political Economics of Development Clusters*. Oxford: Princeton University Press.

Besley, Timothy, and Torsten Persson. 2013. "Taxation and Development." In *Handbook of Public Economics*, edited by Alan Auerbach, Raj Chetty, Martin Feldstein, and Emmanuel Saez, 51–110. Amsterdam: North Holland.

Bolt, Jutta, Robert Inklaar, Herman de Jong, and Jan Luiten van Zanden. 2018. "Rebasing 'Maddison': New Income Comparisons and the Shape of Long-Run Economic Development." Maddison Project Working Paper 10.

Brautigam, Deborah, Odd-Helge Fjeldstad, and Mick Moore, eds. 2008. *Taxation and State-Building in Developing Countries: Capacity and Consent*. Cambridge: Cambridge University Press.

Central Bureau of Health Intelligence (CBHI). 2019. *National Health Profile 2019*. New Delhi: Ministry of Health and Family Welfare.

Centre for the Study of Developing Societies (CSDS). 2015. "Trust in Institutions." In *Democracy in India: A Citizens' Perspective,* 53–74. Delhi: Lokniti.

Centre for Tax Policy and Administration. 2011. *Tax Administration in OECD and Selected Non-OECD Countries: Comparative Information Series (2010)*. Paris: OECD.

Constituent Assembly Debates. 1948. "Constituent Assembly of India Debates (Proceedings)—Volume VII." New Delhi, India, November 4, 1948. https://www.constitutionofindia.net/constitution_assembly_debates/volume/7/%C2%AD1948-11-04.

Dahlström, Carl, Victor Lapuente, and Jan Teorell. 2012. "The Merit of Meritocratization: Politics, Bureaucracy, and the Institutional Deterrents of Corruption." *Political Research Quarterly* 65 (3): 656–68.

Das, Sabyasachi, and Gaurav Sabharwal. 2017. "Whom Are You Doing a Favor to? Political Alignment and Allocation of Public Servants." https://dassabyasachi.files.wordpress.com/2014/05/alignment_paper_april2017_v4.pdf (accessed April 3, 2019).

Dasgupta, Aditya, and Devesh Kapur. 2019. "The Political Economy of Bureaucratic Overload: Evidence from Rural Development Officials in India." Unpublished.

Dreze, Jean, and Amartya Sen. 1989. *Hunger and Public Action*. Oxford: Clarendon Press.

Finan, Frederico, Benjamin A. Olken, and Rohini Pande. 2017. "The Personnel Economics of the Developing State." In *Handbook of Economic Field Experiments*, edited by Abhijit Vinayak Banerjee and Esther Duflo, 467–514. Amsterdam: North Holland.

Fisman, Raymond, Florian Schulz, and Vikrant Vig. 2014. "The Private Returns to Public Office." *Journal of Political Economy* 122 (4): 806–62.

Government of India (GOI). 2015. *Report of the Seventh Central Pay Commission*. New Delhi: Government of India.

Gupta, Aashish, Sangita Vyas, Payal Hathi, Nazar Khalid, Nikhil Srivastav, Dean Spears, and Diane Coffey. 2019. "Persistence of Solid Fuel Use Despite Increases in LPG Ownership: New Survey Evidence from Rural North India." https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&ved=2ahUKEwi64obzqeXlAhUkwFkKHUVaBJAQFjAEegQIAxAC&url=https%3A%2F%2Friceinstitute.org%2Fwp-content%2Fthemes%2Frice%2Fdownloadpdf.php%3Fpfile%3Dhttps%3A%2F%2Friceinstitute.org%2Fwp-content%2Fuploads%2F2019%2F03%2Fgupta-vyas-et-al-2019-persistence-of-solid-fuel-use-in-rural-north-india.pdf&usg=AOvVaw1jI52d2bxur7aDjBtzIsln (accessed April 3, 2019).

Hanna, Rema, and Shing-Yi Wang. 2017. "Dishonesty and Selection into Public Service: Evidence from India." *American Economic Journal: Economic Policy* 9 (3): 262–90.

Hirschman, Albert O. 1967. *Development Projects Observed*. Washington, DC: Brookings Institution.

Hirschman, Albert O. 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge: Harvard University Press.

India Census Commissioner. 1951. *Census of India*. New Delhi: India Census Commissioner.

Iyer, Lakshmi, and Anandi Mani. 2012. "Traveling Agents: Political Change and Bureaucratic Turnover in India." *Review of Economics and Statistics* 94 (3): 723–39.

Joshi, Vijay, and I. M. D. Little. 1994. *India: Macroeconomics and Political Economy, 1964–1991*. Washington, DC: World Bank.

Kapur, Devesh, and Madhav Khosla, eds. 2019. *Regulation in India: Design, Capacity, Performance*. Oxford: Hart Publishing.

Keefer, Philip, and Stuti Khemani. 2004. "Why Do the Poor Receive Poor Services?" *Economic and Political Weekly* 39 (9): 935–43.

Khosla, Madhav. Forthcoming. *India's Founding Moment: The Constitution of a Most Surprising Democracy*. Cambridge: Harvard University Press.

Lehne, Jonathan, Jacob N. Shapiro, and Oliver Vanden Eynde. 2018. "Building Connections: Political Corruption and Road Construction in India." *Journal of Development Economics* 131: 62–78.

Mani, Anandi, and Sharun Mukand. 2007. "Democracy, Visibility and Public Good Provision." *Journal of Development Economics* 83 (2): 506–29.

Mehta, Pratap Bhanu. 2007. "India's Unlikely Democracy: The Rise of Judicial Sovereignty." *Journal of Democracy* 18 (2): 70–83.

Ministry of Finance. 2018. *Economic Survey, 2017–18*. New Delhi: Government of India.

Ministry of Finance. 2016. *Economic Survey, 2015–16*. New Delhi: Government of India.

Ministry of Finance. 2012. *Economic Survey, 2011–12*. New Delhi: Government of India.

Ministry of Finance. 2002. *Economic Survey, 2001–02*. New Delhi: Government of India.

Ministry of Finance. 1996. *Economic Survey, 1995–96*. New Delhi: Government of India.

Ministry of Finance. 1985. *Economic Survey, 1984–85*. New Delhi: Government of India.

Ministry of Finance. 1975. *Economic Survey, 1974–75*. New Delhi: Government of India.

Ministry of Finance. 1973. *Economic Survey, 1972–73*. New Delhi: Government of India.

Ministry of Health and Family Welfare. 2017. *India Fact Sheet 2015–16*. Mumbai: International Institute for Population Sciences.

Ministry of Heavy Industries and Public Enterprises. 2018. *Public Enterprises Survey 2016–17*. New Delhi: Government of India.

Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mohpal. 2017. "The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India." *Journal of Public Economics* 145: 116–35.

National Sample Survey Office. 2015. *Key Indicators of Social Consumption in India: Education (NSS 71st Round)*. New Delhi: Ministry of Statistics and Programme Implementation.

Niehaus, Paul, and Sandip Sukhtankar. 2013. "Corruption Dynamics: The Golden Goose Effect." *American Economic Journal: Economic Policy* 5 (4): 230–69.

Page, Lucy, and Rohini Pande. 2018. "Ending Global Poverty: Why Money Isn't Enough." *Journal of Economic Perspectives* 32 (4): 173–200.

Persson, Anna, and Bo Rothstein. 2015. "It's My Money: Why Big Government May Be Good Government." *Comparative Politics* 47 (2): 231–49.

Polio Global Eradication Initiative. 2016. *Polio in India: Fact Sheet*. Geneva: Polio Global Eradication Initiative.

Princeton Review. n.d. "United States Military Academy." https://www.princetonreview.com/college/united-states-military-academy-1023919.

Pritchett, Lant. 2009. "A Review of Edward Luce's 'In Spite of the Gods: The Strange Rise of Modern India.'" *Journal of Economic Literature* 47 (3): 771–80.

Przeworski, Adam, Michael E. Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990*. Cambridge: Cambridge University Press.

Purohit, Bhaskar, Tim Martineau, and Kabir Sheikh. 2016. "Opening the Black Box of Transfer Systems in Public Sector Health Services in a Western State in India." *BMC Health Services Research* 16 (1): 419.

Quraishi, S. Y. 2019. "What It Takes to Run an Election for India." *New York Times*, April 25. https://www.nytimes.com/2019/04/25/opinion/india-elections.html.

Radkar, Anjali. 2012. "Risk of Maternal Mortality: Indian Scenario." In *Global Health: A Challenge for Interdisciplinary Research*, edited by Martin Kappas, Uwe Groß, and Dermot Kelleher, 117–32. Göttingen: Universitätsverlag Göttingen.

Ramachandran, Vimala, Tara Béteille, Toby Linden, Sangeeta Dey, Sangeeta Goyal, and Prerna Goel Chatterjee. 2018. *Getting the Right Teachers into the Right Schools: Managing India's Teacher Workforce*. Washington, DC: World Bank.

Rauch, James E., and Peter B. Evans. 2000. "Bureaucratic Structure and Bureaucratic Performance in

Less Developed Countries." *Journal of Public Economics* 75 (1): 49–71.

Ren, Xuefei. 2015. "City Power and Urban Fiscal Crises: The USA, China, and India." *International Journal of Urban Sciences* 19 (1): 73–81.

Reserve Bank of India. 2018. "Basic Statistical Returns of Scheduled Commercial Banks in India." https://www.rbi.org.in/Scripts/AnnualPublications.aspx?head=Basic+Statistical+Returns.

Rodrik, Dani, and Arvind Subramanian. 2003. "The Primacy of Institutions (and What This Does and Does Not Mean)." *Finance and Development* 40 (2): 31–34.

Roy, Tirthankar. 1996. "The Role of the State in Initiating Development: A Study of Interwar South and Southeast Asia." *Indian Economic and Social History Review* 33 (4): 373–401.

Roy, Tirthankar. 2011. *The Economic History of India, 1857–1947*. Oxford: Oxford University Press.

Schiavo-Campo, Salvatore, Giulio de Tommaso, and Amitabha Mukherjee. 1997. "Government Employment and Pay: A Global and Regional Perspective." Policy Research Working Paper 1771.

Sheahan, Megan, Yanyan Liu, Christopher B. Barrett, and Sudha Narayanan. 2018. "Preferential Resource Spending under an Employment Guarantee: The Political Economy of MGNREGS in Andhra Pradesh." *World Bank Economic Review* 32 (3): 551–69.

Sukhtankar, Sandip, and Milan Vaishnav. 2015. "Corruption in India: Bridging Research Evidence and Policy Options." *India Policy Forum* 11: 193–276.

Union Public Service Commission UPSC. 2014–2016. "Annual Reports." https://www.upsc.gov.in/annual-reports.

United Nations Office on Drugs and Crime (UNODC). 2019. *Global Study on Homicide 2019*. Vienna: UNODC.

US Office of Personnel Management. 1940–2014. "Historical Federal Workforce Tables: Executive Branch Civilian Employment since 1940." https://www.opm.gov/policy-data-oversight/data-analysis-documentation/federal-employment-reports/historical-tables/executive-branch-civilian-employment-since-1940/.

US Postal Service. 2014. *2014 Annual Report to Congress*. Washington, DC: USPS.

Wade, Robert. 1982. "The System of Administrative and Political Corruption: Canal Irrigation in South India." *Journal of Development Studies* 18 (3): 287–328.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 55–74

The Great Indian Demonetization

Amartya Lahiri

On November 8, 2016, the Prime Minister of India, Narendra Modi, took the nation by surprise by announcing that the government was demonetizing currency with denominations of 500 or 1,000 rupees, with immediate effect. This amounted to the demonetization of 86 percent of the Indian currency in circulation. Holders of the demonetized currency were given till December 31, 2016 to exchange their demonetized bills for newly issued currency, which would be in denominations of 500 and 2,000 rupees.

Modi gave two main reasons for the move: first, it would allow the state to seize the wealth in the economy that was accumulated through undeclared income. Hence, this was to be a decisive blow against corruption. Second, it would eliminate the scourge of counterfeit currency that was circulating in the economy. This second motive, while laudable, seemed aimed at a small target because estimates from the Indian Statistical Institute suggested that counterfeit currency accounted for a bare 0.025 percent of the currency in circulation (as reported in Chauhan 2016).

In subsequent days, two other motives were added to the narrative. Third, demonetization was intended to be a way of pushing India toward a modern digitized economy, which would be less reliant on cash. More digitized payments would bring a larger share of the informal Indian economy into the organized and formal sector. Fourth, by forcing people to convert their old cash into the new currency through the banking system, it was both bringing unaccounted money into the

■ *Amartya Lahiri is Royal Bank Faculty Research Professor, Vancouver School of Economics, University of British Columbia, Vancouver, British Columbia, Canada. His email address is amartya.lahiri@ubc.ca.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.55.

formal tax network and generating greater digital footprints to track individuals and firms who were hitherto hidden from the tax network.

After an extended counting process, when the dust cleared, the Reserve Bank of India announced that over 99 percent of the demonetized currency had been returned to it through the commercial banks. Also, within a year of the demonetization, currency in circulation in the economy was also back to its predemonetization level.

Panel A of Figure 1 shows the time paths of three different measures of money: M0, M1, and M2. The units are in millions of rupees. M0 measures currency in circulation, plus deposits by bankers and others with the Reserve Bank of India. M1 includes currency, demand deposits with the banking system, and other deposits with the RBI. M2 adds savings deposits of post office savings banks to M1. As can be seen from the figure, by the end of March 2017, both M1 and M2 were just 2.1 and 2.9 percent below their October 2016 levels. M0, on the other hand, remained 15 percent below its predemonetization level. In fact, it wasn't until January 2018 that M0 recovered to its predemonetization level.

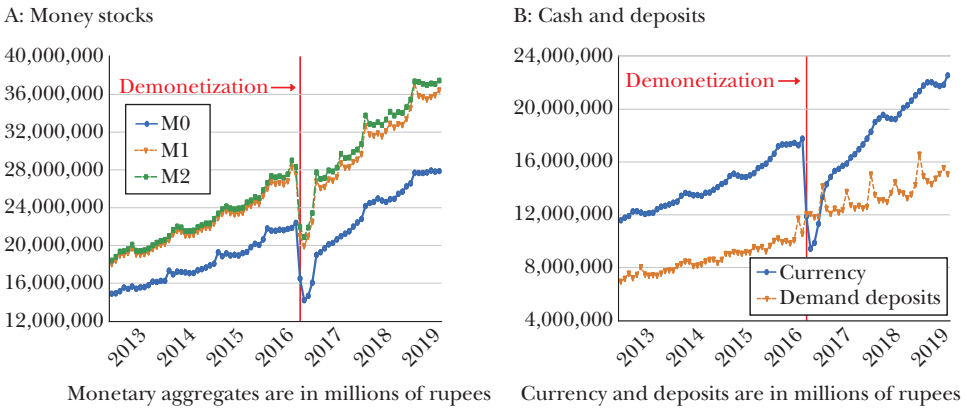
There were two ways in which the Indian public could exchange the demonetized cash: they could either swap the old currency for new currency (subject to daily limits), or they could deposit the old cash in their bank accounts. Panel B of Figure 1 shows the contrasting behavior of currency in circulation and bank deposits (which comprise saving and checking deposits) during the episode. Currency in circulation fell by around 8.4 trillion rupees while bank deposits rose by a meager 1.5 trillion rupees between October (the last month before demonetization) and December 31, 2016 (the last date for exchanging the old bills for new ones). Most of the demonetized currency was instead deposited in time deposits, which rose by over 4 trillion rupees during this period.

The banks, in turn, parked the returned cash with the Reserve Bank of India first in the form of bankers' deposits and subsequently in special purpose bonds issued by the RBI. Since most of the demonetized currency was eventually returned, the overall level of RBI liabilities barely changed during the entire episode.

While the move was initially hailed as courageous and transformative by some commentators, the mood rapidly gave ground to widespread concerns regarding: (1) the preparedness of the Reserve Bank of India to manage the process of remonetizing the economy, (2) the potential of demonetization to achieve the stated goals, (3) and the costs of the move for the Indian economy. With two years having passed since the enactment of the policy, what does the evidence suggest about the effects of India's demonetization?

The evidence points to demonetization having mostly failed to have achieved its stated objectives. The goal of eradicating black wealth and corruption by demonetizing currency was problematic from the start, given the widespread acknowledgement of the fact that undeclared income is seldom held for long periods in terms of cash. Moreover, demonetizing currency, which attacks a stock, does little to impede the fresh creation of undeclared income, which is a flow

Figure 1
Demonetization and Money Stocks



Source: The data comes from Reserve Bank of India (2019).
Note: The left panel shows the path of three different measures of money during the period January 2013–September 2019. The units are in millions of rupees and the frequency monthly. M0 measures currency in circulation, plus deposits by bankers and others with the Reserve Bank of India (RBI). M1 includes currency, demand deposits with the banking system, and other deposits with the RBI. M2 adds savings deposits of post office savings banks to M1. The right panel shows the monthly data on currency in circulation and demand deposits in the banking system in millions of rupees during January 2013–September 2019.

problem. The second goal of destroying counterfeit currency was suspect to start with given the very low-estimated counterfeit currency in circulation in India.

An examination of the growth in digitized payments, in the tax base and in tax revenues, suggests that the move achieved little in terms of changing the predemonetization trends in these measures. Digitized payments were growing exponentially in India prior to 2016, and they have continued on the same nonlinear trend. I also do not detect any systematic impact of demonetization on either the number of tax filers or tax revenues. Of course, given the relatively short period of time since the demonetization, these conclusions on the time trends in digital transactions and taxes should be viewed as tentative. On the cost side, however, there appears to be strong evidence that demonetization reduced output and employment, especially in the informal sector. However, these losses were likely temporary rather than being permanent. On balance, demonetization appears to have failed the cost-benefit analysis of public policy initiatives: it had little success in achieving its stated goals while having imposed significant costs on the public.

In the next section, I place India's demonetization initiative in context by describing the intellectual arguments for demonetization as well as the experience of two other demonetization exercises that were carried out in the past in India. I then examine the preparedness of the central bank in dealing with the mechanics

of demonetization. The discussion then presents the evidence in the context of the logic of the stated goals and considers evidence on the costs of demonetization before offering a brief conclusion.

Intellectual and Historical Context

Demonetization as a tool for fighting crime, tax evasion, and activities in the underground economy has been advocated in the past. One of the more well-known recent contributions along these lines was made by Ken Rogoff (2016, 2017). The argument rests on the premise that, in an international context, many underground economy activities are financed using large-denomination currency notes. Following World War II, Britain and other European countries fought back against illicit wartime speculative wealth gains by demonetizing high denomination bills. In 1969, the United States demonetized bills with denominations of $500 and higher; in 2017, the European Central Bank demonetized the 500-euro bill.

A unique aspect of the Indian measure was that it was carried out during a period of economic stability, but with very little time given to the public to exchange their demonetized bills. This created the potential for a lot of disruption and inconvenience since the demonetized bills, especially the 500-rupee bill (worth about US$14 at prevailing exchange rates), were heavily in use for daily transactions.

The demonetization of 2016 was not the first such episode in Indian monetary history either. There were two other episodes in the post-World War II era with remarkably similar underlying justifications: one in 1946, the other in 1978.

Soon after the end of World War II, on January 12, 1946, the Government of India demonetized all currency bills of denomination 500 rupees and above. In the lead-up to that decision, the finance member of the Governor General of India's Executive Council, Sir Archibald Rowlands, cited the Bank of England's decision to demonetize currency after the war "as one more concrete example for the Indian government to follow in its fight against black market money and tax evasions which have now assumed enormous proportions."[1] There were of course officials who were skeptical of the effectiveness of measure at the time, including the then-Governor and Deputy Governor of the Reserve Bank of India. When all the exchanges were done, it turned out that 94 percent of the demonetized currency was returned to the RBI. The scheme was generally regarded as a failure because not much was garnered in the form of unreturned currency, while the demonetization caused considerable hardship to the general public. Moreover, the higher denomination bills were all reintroduced by 1954.

The second such episode was in 1978. On January 16, 1978, the government demonetized all currency bills of denominations 1,000 rupees and above. In contrast to the 2016 measure, which demonetized 86 percent of the currency

[1] Facts and the background surrounding this episode can be found on page 706 of volume 1 of the fascinating history of the Reserve Bank of India (1970).

in circulation, the 1978 measure only affected approximately 1.5 percent of the currency. As a result, the disruption for the general public was limited. This measure was also opposed by the governor of the Reserve Bank of India at the time, I.G. Patel. Amongst other reservations, Patel (2002) held that "such an exercise seldom produces striking results" and "the idea that black money or wealth is held in the form of notes tucked away in suit cases or pillow cases is naïve." The move was marginally more successfully than the 1946 experience in that 86 percent of the demonetized currency was exchanged for lower denomination bills.

The remarkable part about the 1946 and 1978 episodes was the similarity of the motivation behind them as well as the concerns regarding their efficacy in achieving the stated objectives. In addition, the two previous episodes were similar in that most of the demonetized currency was successfully converted by the public. This rendered the objective of taxing undeclared income unfulfilled for the most part.

The Preparedness of the Reserve Bank of India

Prime Minister Modi announced the demonetization of 500- and 1,000-rupee currency bills on November 8, 2016. It was later revealed that the Board of the Reserve Bank of India had met earlier that evening to consider a letter from the Ministry of Finance that the Government of India received the previous day, along with a memorandum from a deputy governor recommending the demonetizing. Two key reasons for the proposal cited in the government letter were: (1) between 2011 and 2016, the supply of 500- and 1,000-rupee bills had grown by 76 and 108 percent, respectively, while India's economy had only grown by 30 percent during this period; and (2) cash typically facilitated "black money." The board was further told that the measure was also intended to encourage greater financial inclusion and to incentivize greater digitization of the economy.

The board approved the proposal, but not before making a few trenchant comments. It noted that the measure may not have the desired effect on black money because most people do not hold undeclared wealth in cash. It further worried about the negative effects on growth that were likely to occur in the short run. Possibly the most damning observation was that the primary fact on which the government had based its proposal—that the supply of 500- and 1,000-rupee bills had far outstripped the growth rate of the economy—was simply wrong. The board pointed out the embarrassing fact that the government had compared GDP growth in real terms with the growth of currency supply in nominal terms. In fact, nominal GDP growth had summed to over 80 percent between 2011 and 2016 and hence was in line with the growth of the currency bills to be demonetized.[2]

The minutes suggest that the board was assured that demonetization had been under discussion between the Reserve Bank of India and the government

[2] See the Minutes of the Five Hundred and Sixty-First Meeting of the Central Board of Directors of the Reserve Bank of India (Reserve Bank of India 2016).

for the preceding six months, during which these issues had been considered. The ex-Governor of the Reserve Bank of India, Raghuram Rajan, whose term as governor had ended on August 31, 2016, has gone on record confirming this. He said that the RBI had indeed been consulted about demonetization and had advised the government against it (as reported in *Hindu Business Line* 2018).

The preparation of the Reserve Bank of India for this massive operation came into severe focus almost immediately as automatic teller machines ran out of cash for long periods of time across the length and breadth of the country, including the major metropolitan cities. Moreover, when the automatic teller machines had supplies of the new currency, most of it, at least initially, was in the form of 2,000-rupee denomination bills, which was not helpful for daily transactions whose average cash value tended to be much smaller. The process of remonetizing the economy with the new currency bills proved to be slow and severely disruptive for regular commercial transactions.

A further source of concern regarding the preparedness of the Reserve Bank of India for a policy measure of this scale came in the form of the multiple circulars that it issued after the initial notifications announcing the demonetization. The RBI issued 57 official circulars between November 9 and December 31, 2016, which kept revising the conditions under which the public could make deposits, withdrawals, and exchanges of the demonetized currency. For example, over-the-counter exchange of demonetized currency was initially limited to 4,000 rupees per person per day. This daily limit was first raised to 4,500 rupees and then reduced to 2,000 rupees before being completely stopped starting November 24th. On withdrawals from bank accounts, initially daily over-the-counter cash withdrawals were capped at 10,000 rupees with a weekly limit of 20,000 rupees. This weekly limit was subsequently raised to 24,000 rupees, while the over-the-counter limit of 10,000 rupees was withdrawn. Withdrawals via automatic teller machines were initially restricted to 2,000 rupees per day per card before being raised to 2,500 and then 4,000 rupees per day per card.

The rules governing deposits were also constantly being revised. For customers with updated identity documentations, known as Know-Your-Customer or KYC norms, initially there was no capping on the amount to be credited to the account. For non-KYC compliant account holders, a maximum value of 50,000 rupees of demonetized bills could be deposited. On November 16, 2016, the Reserve Bank of India announced that all cash deposits exceeding 50,000 rupees in value needed to be supplemented with a copy of the taxpayer identification card number (known as PAN card) in case the account did not have that information.

The combination of the slow stocking of automatic teller machines with the new cash, the spate of revised notifications, the limited supply of new 500-rupee bills, and the relative excess of new 2,000-rupee bills, which were less useful for transactions purposes, suggested that the institution that had been tasked with implementing the policy was not adequately prepared. Rather, the policy was thrust upon the Reserve Bank of India, which then scrambled to implement the policy as best as it could.

Achieving the Stated Goals

Amongst the various stated policy goals, three of the early ones were: (1) to seize the black wealth created through undeclared income that was stored in the form of cash holdings, (2) increase the tax base by forcing people to exchange demonetized bills through the banking sector, (3) and to convert the economy into a more digitized one that was less dependent on cash.

Seizing Black Wealth: Direct and Indirect Methods

There are two ways of seizing unaccounted income or black wealth. The first is by taxing it directly, while the second is indirectly by bringing underground economy transactions into the tax net. We examine the effect of demonetization on both of these channels.

For the government to be able to directly seize black (unaccounted) wealth through demonetization, a necessary condition was that the share of demonetized currency that was returned to the Reserve Bank of India be significantly less than 100 percent. Given that over 99 percent of the old cash was returned, this direct method of capturing unaccounted wealth did not work. Nevertheless, in assessing whether the demonetization could have even been expected to achieve this goal, it is useful to conduct a few back-of-the-envelope computations.

Black money has both a stock and a flow aspect. To assess the impact of demonetization, we need estimates for both. In a World Bank study, Schneider, Buehn, and Montenegro (2010) estimate the parallel economy in India to be around 25 percent of GDP. This gives an estimate of the flow share of the underground economy.[3] The wealth share of the underground economy is more difficult. Credit Suisse (2014) estimates the wealth-to-GDP ratio in India to be around two. If wealth creation is similar for both declared and undeclared income, this would suggest that black wealth in India is about 50 percent of GDP. It is likely larger because the saving rate out of undeclared income is probably greater than that out of declared income. Nevertheless, one can use these two estimates for a rough calculation of the amount of black wealth and black income that demonetization could have realistically been expected to mop up.

The demonetized money was about 10 percent of GDP. Even if the entire amount had been left unexchanged, it would have amounted to around 40 percent of the underground economy (or black income) and 20 percent of black wealth.

[3]The Schneider, Buehn, and Montenegro (2010) estimate of the underground economy is an attempt to measure output that is deliberately not reported in order to avoid detection. It is different from the estimated informal economy share of Indian GDP of 45 percent. The estimated informal economy is part of India's official GDP estimates. The estimate for nonagricultural informal sector output is derived from enterprise surveys of unincorporated firms. Estimates of labor value added in the unincorporated sector derived from the enterprise surveys are combined with estimates of labor supply to the informal sector derived from household employment surveys to arrive at the estimate for nonagricultural informal sector output. Estimates for agricultural informal sector output are derived by combining land-use statistics with data on cropping an area by crop and cost of inputs.

Given the historical precedents from 1946 and 1978, above, a reasonable working guess would have been 85–90 percent of the demonetized cash would be exchanged. Hence, the maximum amount that this move could have been expected to garner was around 2–3 percent of the black wealth in India (or 4–6 percent of black income).

These estimates, which would have been easy to compute before enacting the policy, seem rather small given the extent of the disruption to the economy. As it turned out, these gains were close to zero since over 99 percent of the demonetized cash was exchanged by the public. At least on this dimension, the policy seems to have been poorly conceptualized.

The second way in which demonetization could seize unaccounted wealth is indirectly through its effect on the tax base. To see this, note that there were two ways of exchanging old bills: (1) over-the-counter exchanges of old bills for new ones and (2) depositing old bills in one's bank account and withdrawing new cash at a later date. The Reserve Bank of India imposed severe restrictions on the first option by limiting the maximum amounts that could be exchanged over-the-counter at banks. Inasmuch as the public returned the old bills through the second option, depositors would be traceable. Hence, the government could potentially identify individuals/entities whose deposits were higher than the norm. The government could then examine the tax and income footprints of these depositors more closely to identify tax evaders and confiscate some of their unaccounted-for wealth.

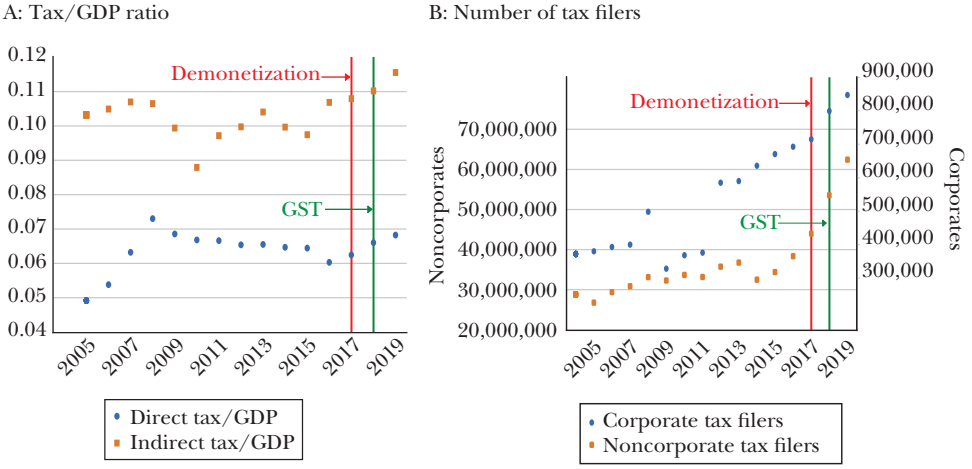
We investigate this indirect effect of demonetization by examining the time series behavior of two different indicators. The first considers the evolution of the tax/GDP ratio in India before and after demonetization, while the second examines the evolution of the number of tax filers before and after demonetization.

Before proceeding further, it is important to note two caveats. First, India enacted a key tax reform in July 2017 when it introduced a Goods and Services Tax (GST). GST replaced a complicated web of disparate indirect tax schemes that varied across states both in magnitude and extent. The GST reform had been in the works for over a decade. As a result of this change, assessing the impact of demonetization on tax revenues accruing to the government is problematic since the two measures occurred in such close proximity. Second, we only have three years of tax data after demonetization. This makes it difficult to draw any definitive econometric conclusions.

In Figure 2, panel A depicts the path of the tax/GDP ratio in India from 2005. The figure plots direct and indirect tax ratios separately. Direct taxes are primarily composed of personal income taxes and corporate taxes. Indirect taxes are comprised of sales taxes, customs duties, and excise duties.[4] The vertical lines on the graphs mark the fiscal years in which demonetization and the Goods and

[4]The tax data are annual. Because the Indian fiscal year goes from April 1 to March 31, the years in the figures refer to the fiscal year. Thus, 2015 refers to the fiscal year 2014–2015 that ended on March 31, 2015.

Figure 2
Demonetization and Tax Revenues



Source: The tax data is from Government of India (2019), the Central Board of Direct Taxes (CBDT) in India. Data on the number of tax filers comes from Government of India (2019) and Comptroller and Auditor General of India (2019). The GDP data comes from Reserve Bank of India (2019).
Note: The figure in the left panel depicts the tax-GDP ratio for both direct and indirect taxes for the period 2005–2019. The figure in the right panel reports the number of tax filers in India during 2005–2019. The left axis reports the number of noncorporate tax filers, while the right axis reports the number of corporate tax filers.

Services Tax were introduced. Demonetization occurred in fiscal year 2016–2017, while GST happened in fiscal year 2017–2018.

A couple of features are noteworthy. First, direct taxes typically account for just about one-third of overall tax revenues in India. This is due to the very small number of individual and other noncorporate taxpayers in India (around 44 million in 2017, less than 10 percent of the labor force). The abysmal state of direct taxes has been a long-running public finance concern in India. It partly reflects the low income of most of the workforce, but is also symptomatic of widespread tax evasion. Second, there does appear to be a mild increase in both the direct and indirect tax/GDP ratios in 2017 relative to 2016 (the fiscal year before demonetization). However, the figure also shows that both the direct and indirect tax ratios in 2018 were not very different from their past trends. Thus, the direct tax ratio in India has been stable between 6 and 7 percent since 2010. Its levels in 2018 and 2019 were 6.7 and 6.9 percent, respectively. Interestingly, these levels for the direct tax ratio are below the levels it reached in 2008–2009. The indirect tax/GDP ratio has been on a gradually rising path except for declines in 2014 and 2015. Neither demonetization nor Goods and Services Tax appear to have pushed the indirect tax/GDP ratio off its recent trend path.

Based on the limited evidence of three years of post-demonetization tax revenue, it is hard to argue that demonetization induced a sharp increase in the collection of tax revenues. Clearly, a conclusive assessment of the impact of demonetization on tax revenues would require a few more years of data, as well as decoupling the effects of Goods and Services Tax from demonetization.

The tax revenue data do not distinguish between the tax rate and the number of tax filers. One conjectured effect of demonetization was that it would bring more individuals and firms into the tax net by forcing them to exchange their demonetized cash through the formal banking system. Figure 2 examines this hypothesis by plotting the evolution of the number of tax filers in India, broken up by noncorporate and corporate filers. The primary insight from panel B of Figure 2 is that both total and corporate tax filers have been steadily rising since 2014. There doesn't appear to be any sharp increase in the number of tax filers in 2017, which was the year of demonetization. In fact, the figures suggest that there was a sharper increase in the number of tax filers in 2018, which was the year when Goods and Services Tax was introduced, followed by a further increase in 2019. Of course, this could also be the consequence of a delayed response of some tax filers to demonetization.

The general picture that emerges from Figure 2 is that there has been some improvement in public finances in India since 2016, but it is difficult to attribute this to demonetization because the changes appear to be consistent with a prior trend. Hence, the indirect effect of demonetization on seizing undeclared income seems muted at best.[5]

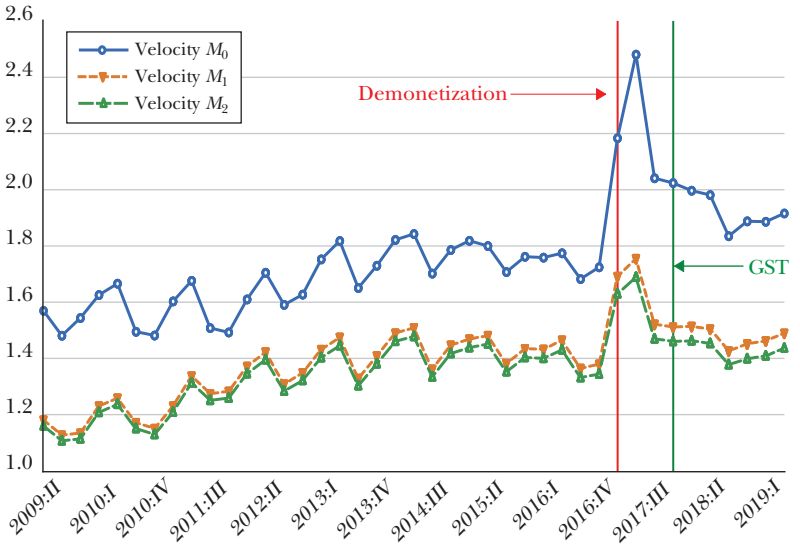
Creating a More Digitized India

The effect of demonetization on the goal of converting India into a more digitized economy is trickier to evaluate. The desire for more digitization originates in the fact that 80 percent of workers, 45 percent of GDP, and a majority of firms in India operate in the informal, unregistered sector. These entities are mostly unregulated and untaxed. Despite the scale of India's economy—1.25 billion people and a labor force of 600 million—the total number of registered individual and noncorporate taxpayers in India, as noted earlier, is a measly 44 million.

This scale of informality in India creates multiple constraints for its economy. First, the small base for direct taxes creates an overdependence on indirect taxation for government revenues, which often results in cascading distortions and efficiency losses. Second, the widespread informal organization of production impedes the penetration of banks and formal finance which, amongst other factors, tends to cause a preponderance of small-scale, low-productivity establishments and firms. In

[5]Another popular method of evaluating the response of taxes is "tax buoyancy," which measures the elasticity of taxes with respect to nominal GDP. An increase in tax buoyancy could thus indicate either an increase in the average tax rate or an increase in the number of people paying taxes. The conjectured effect of demonetization on bringing people into the tax net would typically operate through the second channel. The tax buoyancy numbers in India are so volatile that it is impossible to detect any trend or trend break from it. The tax buoyancy results are available from the author upon request.

Figure 3
Demonetization and the Velocity of Money



Source: The source for data on GDP and monetary aggregates is Reserve Bank of India (2019).
Note: Velocity is calculated as the ratio of nominal GDP to the relevant monetary aggregate. M0 measures currency in circulation, plus deposits by bankers and others with the Reserve Bank of India (RBI). M1 includes currency, demand deposits with the banking system, and other deposits with the RBI. M2 adds savings deposits of post office savings banks to M1.

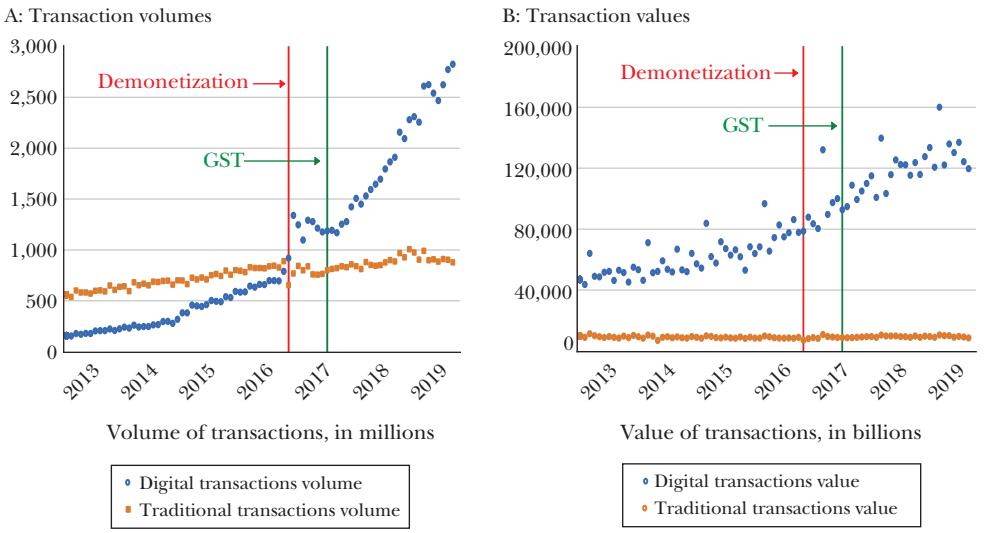
as much as demonetization induces greater digitization of the economy, it would also reduce these constraints.

Clearly, the greater the proportion of transactions that are done through electronic payments such as bank-to-bank money transfers, debit cards, and credit cards (for both business-to-business and business-to-customer transactions), the greater the digital footprints in the economy. How successful has demonetization been in increasing the speed of digitization of the economy?

If demonetization induced the Indian public to switch out of cash transactions, then one should observe a rise in the velocity of money. Velocity of money is defined as the ratio of nominal GDP to the stock of money. It captures the speed with which money circulates in the economy in order to buy the flow of goods being produced. Naturally, the estimated value of velocity depends on the measure of money that one uses. The narrower the measure of money, the larger the measured velocity will be. Our interest here is not in the level of velocity, but rather its movements around and after the time of demonetization.

Figure 3 shows the measured velocity for three different monetary aggregates defined earlier, ranging from the M0, which is the narrowest, to M2, which is the broadest. Two features of the figure are worth noting. First, the biggest increase

Figure 4
Demonetization and Digitization



Source: Data on transactions comes from the Monthly Payment and Settlement Indicators of Reserve Bank of India (2019).
Note: The figure shows the volume and value of monthly traditional and digital transactions in India during January 2013–September 2019. The left panel shows the volume of transactions in millions, while the right panel shows the value of transactions in billions of rupees. Traditional transactions are transactions that involve either paper clearing or card transactions at ATMs, while all other transactions are classified as digital.

in velocity around the demonetization period was for M0, which is the narrowest measure of money. Velocity of M1 rose as well but less than for M0. Movements in the velocity of the broader measures of money were extremely muted by contrast; indeed, measures of velocity associated with broader measures of money than M2 showed almost no change in response to demonetization. Second, after three quarters, all the velocity measures returned to their near-term trend levels. It would appear that initially there was some substitution from cash into other payment methods for transactions in response to the monetary shock. Once things normalized, however, the public returned to their usual usage of cash for transactions purposes.

An alternative approach to measure the effect of demonetization on digitization is to examine directly the time paths of digital transactions in the economy. Figure 4 examines the effect of demonetization on the digitization of the Indian economy by plotting the evolution of digital and traditional transactions, both in terms of volumes and value. Traditional transactions are transactions that involve either paper clearing or card transactions at automatic teller machines, while all other transactions are classified as digital.[6]

[6]Card transactions at automatic teller machines are considered as cash-based transactions and consequently collected under traditional transactions.

In Figure 4, panel A shows the volume of both digital and traditional transactions, while panel B shows the corresponding transaction values. A few features of the transactions data are noteworthy. First, the volume of digital transactions had been steadily growing in India and had almost caught up with the volume of traditional transactions.[7] In fact, the volume of digital transactions had almost caught up with the traditional transactions by October 2016. The demonetization of November 2016 caused the volume of digital transactions volume to shoot up on impact, while simultaneously causing a drop in the volume of traditional transactions. These patterns reversed themselves somewhat in subsequent months so that the traditional transactions volume returned to its predemonetization level. The volume of digital transactions did fall back somewhat from its levels during the demonetization months but, nevertheless, stayed well above its predemonetization level. Indeed, digital transactions have consistently exceeded traditional transactions both in levels and growth rates since 2017. Second, the value of digital transactions have been larger and have also been growing faster than traditional transactions for the past decade. However, demonetization does not appear to have affected the trends or levels of either digital or traditional transactions. In fact, the introduction of the Goods and Services Tax reform also appears to have had no effect on the transactions values.

Because the volume of digital transactions has risen discretely post-demonetization while the value of digital transactions has stayed on its trend path, it appears that demonetization may have induced the public to start using digital payment methods for smaller value transactions relative to the predemonetization period.

Two recent papers investigate the effect of demonetization on digitization more formally. Crouzet, Gupta, and Mezzanotti (2019) examine the evidence to assess whether a large temporary shock to the availability of cash could induce a permanent adoption of electronic payment systems and thus induce digitization. Using data from a digital wallet firm called Paytm, the paper shows that the demonetization shock did induce a permanent increase in digitization. However, this adoption effect was crucially dependent on exposure to the demonetization shock. The Reserve Bank of India distributes currency throughout the country using around 4,000 "currency chests," which are managed by individual bank branches. This research identifies areas further away from currency chest banks as areas that were most exposed to the shock and shows that areas that were more exposed to the shock adopted digital payment methods more aggressively. Moreover, areas that adopted digitization more aggressively were also areas that were more likely to have had higher adoption rates prior to the shock and were also likely to be closer to financial hubs. They interpret these findings as evidence of network effects in the adoption of new technologies.

[7] This statement is subject to the caveat that we do not have independent data on the volume and value of cash transactions in the economy, except for the indirect evidence through the velocity of money that we presented in Figure 3.

In a related paper, Aggarwal, Kulkarni, and Ritadhi (2019) use a difference-in-difference approach to confirm the Crouzet, Gupta, and Mezzanotti (2019) result that areas more exposed to the shock saw a larger increase in digital payments. They then show that adoption of digital payment methods was more muted in districts with more informal workers and rural households. They find that the positive digitization effects were concentrated in districts with fewer rural households and greater shares of salaried workers. Aggarwal, Kulkarni, and Ritadhi (2019) interpret this finding as suggesting that digitization was more likely to occur in response to a negative currency shock in areas that had the requisite infrastructure for digital payments already in place. Relative to the data used by Crouzet, Gupta, and Mezzanotti (2019), Aggarwal, Kulkarni, and Ritadhi (2019) differ along two margins. First, they not only use the location information about the currency chest but also use information about the currency disbursements made by the currency chests. Second, they use proprietary zip code level data on digital payments using debit and credit cards issued by a national vendor called RuPay.

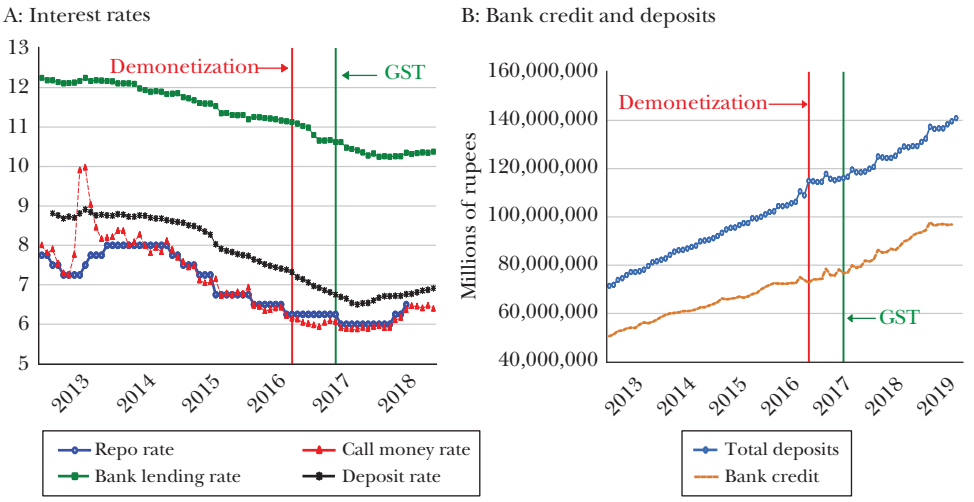
The results of Aggarwal, Kulkarni, and Ritadhi (2019) and Crouzet, Gupta, and Mezzanotti (2019), despite their somewhat different data and methods, point to a common finding. Specifically, the likelihood of demonetization having the desired positive effect on digitization and formalization of the economy depended crucially on the extent of formalization and digitization of the economy already. Put differently, areas that were informal and not very integrated with the formal financial network were unlikely to adopt digitization in response to a shock like demonetization.

Economic Costs of Demonetization

Demonetization clearly upended the daily life of Indians in a significant way. Starting from the immediate constraints faced by individuals and households of conducting daily transactions with a severely diminished supply of cash to the hurdles faced by informal firms trying to pay their suppliers and workers without the standard access to cash, anecdotal evidence abounds on the scale of the disruption. Indeed, newspaper accounts and industry reports at the time highlighted sharp job losses in small and medium manufacturing enterprises as well as a huge increase in the demand for jobs under one of India's biggest rural job guarantee schemes called Mahatma Gandhi National Rural Employment Guarantee Program (MNREGA) (for example, as reported in Nair 2017; Janardhanan 2017).

Estimating the effects of demonetization is difficult because the event is still relatively recent, and as noted earlier, the time series aggregate data are not long enough to allow any credible econometric analysis of the economic consequences. But as an indicator based on current data, Figure 5 shows the path of four different interest rates in India since 2013 as well as the path of bank credit and bank deposits. The interest rates shown in the figure are the repo rate, which is the policy rate of the Reserve Bank of India; the call money rate, which is the rate at which short-term

Figure 5
Interest Rates and Bank Credit Conditions



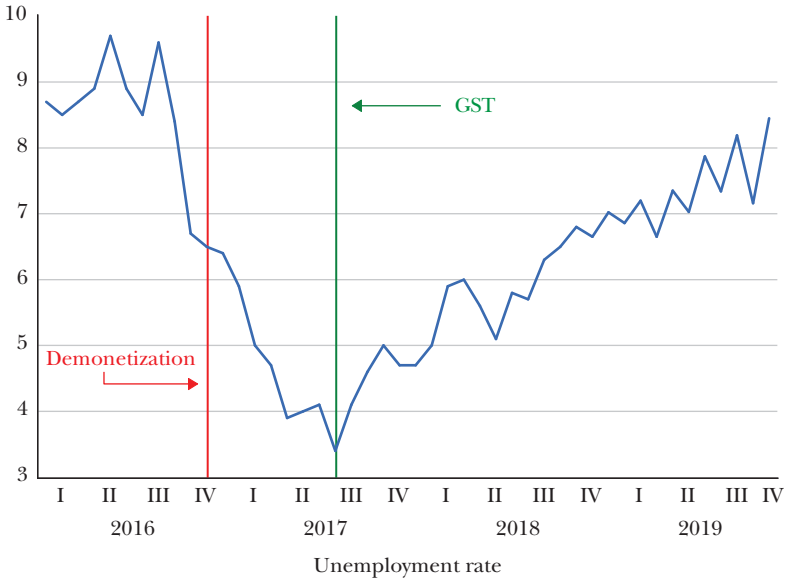
Source: The data on interest rates, deposits, and bank credit comes from the Reserve Bank of India's Database on Indian Economy.
Note: The figure shows the monthly data on different interest rates during January 2013–January 2019 and bank credit and bank deposits in India during January 2013–September 2019. The repo rate is the policy rate of the Reserve Bank of India, and the call money rate is the rate at which short-term funds are borrowed in the overnight money market. The bank lending rate is the weighted average lending rate of banks, while the deposit rate is the weighted average bank-term deposit rate. Bank credit and bank deposits are reported in millions of rupees.

funds are borrowed in the overnight money market; the bank lending rate; and the bank-term deposit rates. All the interest rates other than the repo rate are weighted averages.

Because banks were flush with deposits during the months immediately after the demonetization shock, one might have expected credit conditions to have become significantly easier. However, none of the interest rates showed any sharp movement off their long-run trends around the demonetization date, nor did bank credit pick up in any significant way. In fact, bank credit fell marginally on impact. This is somewhat surprising given that total bank deposits rose by almost 6 trillion rupees on impact of the shock.

The unemployment rate is another variable that one might look at for clues regarding the effects of demonetization. There are no official statistics on unemployment in India currently. However, a private data firm called the Centre for Monitoring the Indian Economy (CMIE) has started collecting high-frequency labor force data since 2016 to fill this gap. The CMIE labor force survey is a longitudinal survey that samples around 160,000 households in three waves every year. Since the surveys are conducted nationally year-round, they publish monthly, quarterly,

Figure 6
Demonetization and Unemployment



Source: The source for unemployment statistics is Center for Monitoring the Indian Economy (2019).
Note: The figure shows the monthly unemployment rate in India for the period January 2016–October 2019. It is based on a nationally representative household survey.

and annual labor force statistics. Figure 6 shows the monthly unemployment rate as reported by the CMIE since 2016.

The figure reveals two interesting features. First, the unemployment rate in India was declining throughout 2016. There is hardly any noticeable effect of demonetization on this declining trend. Second, the unemployment rate begins to rise steeply in India after the introduction of the Goods and Services Tax reform.

While Figure 6 might suggest that demonetization had a tepid effect on unemployment in India, Vyas (2018) presents evidence suggesting that underneath the declining unemployment rate trend though is a steep decline in the labor force that coincides with the demonetization quarter. Using the CMIE monthly and quarterly labor force statistics, Vyas documents two facts. First, relative to the three-month period immediately preceding demonetization (July–October 2016), the number of employed individuals declined by 3.5 million during the period November 2016–February 2017. Second, the CMIE survey also found a dramatic 15 million decline in the size of the labor force between these two periods. Most of this decrease in the labor force was accounted for by a fall in the number of individuals who identified themselves as unemployed. In other words, Vyas suggests that the period of

demonetization coincided with a sharp increase in the number of discouraged workers who simply exited the labor force completely.

Researchers have attempted to get around the limitations of the time series evidence by exploiting the cross-sectional heterogeneity in India. Two recent papers that take this cross-sectional approach to identifying the effects of demonetization are Chodorow-Reich et al. (2018) and Karmakar and Narayanan (2019). Both papers attempt to measure the costs of demonetization by identifying some exogenous cross-sectional variation in exposure to the shock in order to draw causal inference.

Chodorow-Reich et al. (2018) use the variation in remonetization at the currency chests around the districts of India after the demonetization notification as a source of exogenous and random variation. They then measure the cost of demonetization by regressing the cross-sectional outcome variables that vary across districts on the remonetization of the currency chests and other controls. They use a slew of different outcome variables that include "night lights" data, labor force statistics, digitization rates, and others. Based on their estimated cross-sectional responses, they estimate that demonetization induced at least a 2 percentage point decline in GDP in the quarter of demonetization relative to the counterfactual of no-demonetization. They also find that, like the results on digitization described earlier, the output costs of demonetization dissipate over the subsequent months, implying that the effects were transitory.

A potential problem with the identification in Chodorow-Reich et al. (2018) is the assumption that the remonetization at the different currency chests can be treated as exogenous. The validity of the causal inference rests crucially on this identifying assumption. While the paper presents evidence that the rate of distribution of new cash across districts seemed mostly unrelated to variations in the predemonetization levels of different variables, one might nevertheless worry that the distribution of the new cash around the different currency chests may not have been completely random. Indeed, the Reserve Bank of India's (2017) annual report suggests that the distribution of new currency followed a prior plan. Moreover, it would be realistic to expect that the RBI responded to incoming status reports in choosing the allocations of the freshly minted currency during the 52 days between November 9 and December 31, 2016, when the currency exchange was permitted.

The work by Karmakar and Narayanan (2019) uses an alternative identification scheme. They look at a panel dataset on Indian households with information on their asset holdings as well as a host of other indicators such as income, consumption, and others. Their identification scheme is to contrast the response of households who did not have bank accounts on the date of demonetization versus those that did have them. The assumptions underlying this is twofold. First, having a bank account before the demonetization shock was clearly exogenous to demonetization. Second, the real effects of the shock would likely operate through the transactions value of cash. Because access to the new currency was much easier if one had a bank account, the two assumptions jointly imply that those with bank accounts would have smaller disruptions than those without.

The principal findings of Karmakar and Narayanan (2019) are that in December 2016 (the month immediately following the demonetization shock), the 17 percent of households that didn't have bank accounts experienced 2 to 7 percent lower consumption than the control group of households with bank accounts, with the size of the effect varying by the initial asset levels of the household. Moreover, they also found that households without bank accounts tried to find alternative sources of borrowing from various sources at higher rates relative to households that had bank accounts.

Conclusion

The demonetization of 86 percent of the outstanding currency in circulation by the Government of India announced on November 8, 2016, was arguably one of the largest monetary shocks to ever hit the Indian economy. At the end of the exercise, over 99 percent of the demonetized currency was successfully returned by the public in exchange for either new currency bills or claims to new currency. During the transition, however, the demonetization caused almost two months of acute disruption of basic economic activity in a country heavily dependent on cash transactions.

The effect of demonetization in terms of its stated goals were limited at best. Because almost all the demonetized currency was returned to the central bank, it failed in its goal of taxing undeclared income and black (undeclared) wealth. Moreover, available estimates of the circulation of counterfeit currency at the time of demonetization suggested that it was minuscule to start with. Relative to past trends, demonetization does not appear to have had any significant effect on the tax base. There does, however, appear to have been a positive, albeit muted, permanent increase in the degree of digitization of the economy. These conclusions though should be viewed as tentative given that we only have three years of data post-demonetization.

The costs of demonetization are difficult to estimate. However, there are clues. As an example, the large increase in bank deposits during the demonetization period caused a surplus of loanable funds. However, there was almost no impact of this either on the amount of bank loans or in the average lending rate. This tends to suggest that the economic disruption induced by demonetization may have caused a deterioration in the perceived creditworthiness of the average borrower.

Existing research on estimating the costs of demonetization using disaggregated data suggests that it could have lowered output by as much as 2 percentage points during the demonetization quarter. Almost all work in this area also suggests that the costs were temporary and lasted at most two quarters. This is not a surprise because the monetary shock was temporary and the remonetization of the economy was complete in less than two quarters. Available labor market statistics suggest that up to 3.5 million jobs may have been lost during the three months following demonetization while 15 million people may have exited the labor force.

It is surprising, however, that the aggregate statistics do not reveal much effect of the demonetization shock. Perhaps the most striking is the official aggregate

GDP statistic for fiscal year 2016–2017. On January 31, 2019, India's Central Statistical Organization released a revised GDP series, which estimates real GDP growth in the fiscal year 2016–2017 to have been 8.2 percent, the highest since 2011–2012. This implies that India's annual GDP growth increased by 20 basis points in the year of demonetization, relative to the previous year. It is possible that growth in the nondemonetization quarters, particularly the period April–September 2016, saw very rapid economic growth that was partially undone by the negative effects of demonetization during the rest of the year. On the face of it, however, the dissonance between the available cost estimates of demonetization from the disaggregated studies and the estimated increase in aggregate GDP growth from the official statistics for that year represents a puzzle which requires a closer examination.[8]

Demonetization probably had some ancillary effects as well. For example, fighting elections in India requires cash, and there was a major election in the most populous Indian state, Uttar Pradesh, scheduled for February 2017. Demonetization almost surely would have affected the parties that were fighting the Uttar Pradesh election, though the extent of it would likely vary across national parties and regional parties. Little research exists on the political economy dimension of demonetization, and it would certainly be a worthwhile area of future research.

Another issue of importance is the distributional impact of demonetization. Demonetization was packaged as a measure against relatively wealthy individuals who had accumulated undeclared wealth. However, anecdotal accounts suggest that it may instead have disproportionately affected the relatively poorer households working in the informal sector. As more disaggregated household survey data become available over the next few years, this would be another interesting issue to study.

More generally, the Indian experience suggests that demonetization is likely to have a better chance of achieving the goals of fighting crime and tax evasion if larger denomination bills are demonetized. In India, the 500-rupee bills were heavily used for daily transactions. Arguably, the disruptive effects of demonetization would have been more limited if the government had demonetized just the 1,000-rupee bills. Governments contemplating such moves in the future may be better advised to demonetize large denominational bills rather than those that are heavily used for daily transactions.

■ *I would like to thank Paul Beaudry, Viktoria Hnatkovska, Tarun Ramadorai, and the editors of this journal for helpful comments. Special thanks to Timothy Taylor for detailed comments and suggestions. Thanks also to Sujan Bandyopadhyay and Sudipta Ghosh for excellent research assistance. The opinions expressed here are mine and do not reflect the opinions of any institution.*

[8] Intriguingly, the older data prior to the data revision showed a 1.1 percentage point reduction in the annual growth rate in 2016–2017 relative to the previous fiscal year, which was more in line with the disaggregated data. A description of the revision is reported in *Times of India* (2019).

References

Aggarwal, Bhavya, Nirupama Kulkarni, and S. K. Ritadhi. 2019. "Cash Supply Shock and Formalization: Evidence from India's Demonetization Episode." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418430.

Center for Monitoring the Indian Economy. 2019. "Unemployment Rate." Center for Monitoring the Indian Economy. https://www.cmie.com/.

Chauhan, Neeraj. 2016. "Fake Notes Worth Rs 400 Crores in Circulation." *Times of India*, May 11. https://timesofindia.indiatimes.com/india/Fake-notes-worth-Rs-400-crores-in-circulation/articleshow/52214965.cms.

Chodorow-Reich, Gabriel, Gita Gopinath, Prachi Mishra, and Abhinav Narayanan. 2018. "Cash and the Economy: Evidence from India's Demonetization." NBER Working Paper 25370.

Comptroller and Auditor General of India. 2019. *Compliance Audit of Union Government*. New Delhi: Comptroller and Auditor General of India.

Credit Suisse. 2014. *Global Wealth Databook 2014*. Zürich: Credit Suisse.

Crouzet, Nicolas, Apoorv Gupta, and Filippo Mezzanotti. 2019. "Shocks and Technology Adoption: Evidence from Electronic Payment Systems." https://www.kellogg.northwestern.edu/faculty/crouzet/html/papers/TechAdoption_latest.pdf.

Government of India. 2019. "Income Tax Department Time Series Data Nancial Year 2000–01 to 2018–19." Government of India's Income Tax Department. https://dbie.rbi.org.in/DBIE/dbie.rbi?site=home.

Hindu Business Line. 2018. "Demonetization Was 'Not a Good Idea': Rajan." *Hindu Business Line*, April 12. https://www.thehindubusinessline.com/economy/implementation-of-gst-not-unfixable-problem-raghuram-rajan/article23510689.ece.

Janardhanan, Arun. 2017. "Manufacturing Sector Suffers from Considerable Job Loss Post Note Ban." *Indian Express*, January 17. https://indianexpress.com/article/business/economy/demonetisation-manufacturing-sector-suffers-from-considerable-job-loss-post-note-ban-4477687/.

Karmakar, Sudipto, and Abhinav Narayanan. 2019. "Do Households Care about Cash? Exploring the Heterogeneous Effects of India's Demonetization." Research in Economics and Mathematics (REM) Working Paper 073-2019.

Nair, Shalini. 2017. "As Rural Hands Return, NREGA Demand Spikes over 60 per Cent." *Indian Express*, January 9. https://indianexpress.com/article/india/as-rural-hands-return-nrega-demand-spikes-over-60-per-cent-4465577/.

Patel, I. G. 2002. *Glimpses of Indian Economic Policy: An Insider's View*. Oxford: Oxford University Press.

Reserve Bank of India. 1970. *History of the Reserve Bank of India (1935–51)*. Vol. 1. Bombay: Reserve Bank of India.

Reserve Bank of India. 2016. "Minutes of Five Hundred and Sixty First Meeting of the Central Board of Directors of the Reserve Bank of India." December 2016. http://www.humanrightsinitiative.org/download/DeMon%201stattachment.pdf.

Reserve Bank of India. 2017. *Reserve Bank of India Annual Report 2016–17*. Mumbai: Reserve Bank of India.

Reserve Bank of India. 2019. "Database on Indian Economy." 2019. RBI's Database on Indian Economy. https://dbie.rbi.org.in/DBIE/dbie.rbi?site=home.

Rogoff, Kenneth S. 2016. *The Curse of Cash*. Princeton: Princeton University Press.

Rogoff, Kenneth. 2017. "Dealing with Monetary Paralysis at the Zero Bound." *Journal of Economic Perspectives* 31 (3): 47–66.

Schneider, Friedrich, Andreas Buehn, and Claudio E. Montenegro. 2010. "Shadow Economies All over the World: New Estimates for 162 Countries from 1999 to 2007." Policy Research Working Paper 5356.

Times of India. 2019. "Revised GDP Data Shows Year of Demonetisation Was Best for Narendra Modi Government." February 1, https://timesofindia.indiatimes.com/business/india-business/revised-gdp-data-shows-year-of-demonetisation-was-best-for-narendra-modi-government/articleshow/67782001.cms?.

Vyas, Mahesh. 2018. "Using Fast Frequency Household Survey Data to Estimate the Impact of Demonetisation on Employment." *Review of Market Integration* 10 (3): 159–83.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 75–93

Asylum Migration to the Developed World: Persecution, Incentives, and Policy

Timothy J. Hatton

Who is a refugee? The most widely used definition is given by Article 1 of the 1951 Refugee Convention: a person who "owing to well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country…" This definition has been widened in certain treaties—including the 1969 Convention of the Organization of African Unity, the 1984 Cartagena Declaration (in Latin America), and the European Union's 2004 and 2011 Qualification Directives—to include those suffering from persecution on other grounds and those fleeing generalized violence such as war or armed insurrection.

Refugee policy differs from regular immigration policy in two respects. First, in high-income countries, the immigration stream focuses on two groups: individuals with family ties to the receiving country (as is common in Italy, Spain, Japan, Israel, and the United States), or individuals deemed to meet specific labor market criteria (for example, the point systems used in Canada and Australia, or the US H-1B visa system). Immigration policies can be interpreted as serving the interests of the host-country population, either specific individuals such as the sponsors of those coming through family reunification, or the wider economy as in the case of skill-selective labor migration. By contrast, refugees are admitted on the grounds of the benefit to *them* of escaping persecution rather than for any direct benefit to the host society or certain members of it. Indeed, the sole criterion of having a "well-founded fear

■ *Timothy J. Hatton is Professor of Economics at the University of Essex, Colchester, Essex, United Kingdom, and at the Australian National University, Canberra, Australia. His email address is hatton@essex.ac.uk.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.75.

of persecution" is specific to the individual refugee and does not depend on the "value" of that person to the host country. Rather, the rationale for the host society of providing a safe haven for refugees is much more indirect: to meet basic humanitarian concerns.

Second, while immigration policy involves characteristics that are relatively straightforward to verify, the definition of a refugee is much more subjective. Assessing the authenticity of applications for asylum requires destination countries to make an individual assessment, often based on inadequate or incomplete evidence (for example, is a particular migrant truly under threat for political or religious beliefs in that person's home country?). Also, it usually involves an evaluation of the situation in origin countries (for example, what is the extent of human rights abuse?). It can be difficult to separate refugees as defined in international agreements from those who wish to migrate for economic reasons. This is because most of the hundreds of thousands who apply for political asylum each year come from countries that are both strife-prone *and* poor, places where suffering genuine fear of persecution is a distinct possibility, and also where the economic gains to emigration would be large.

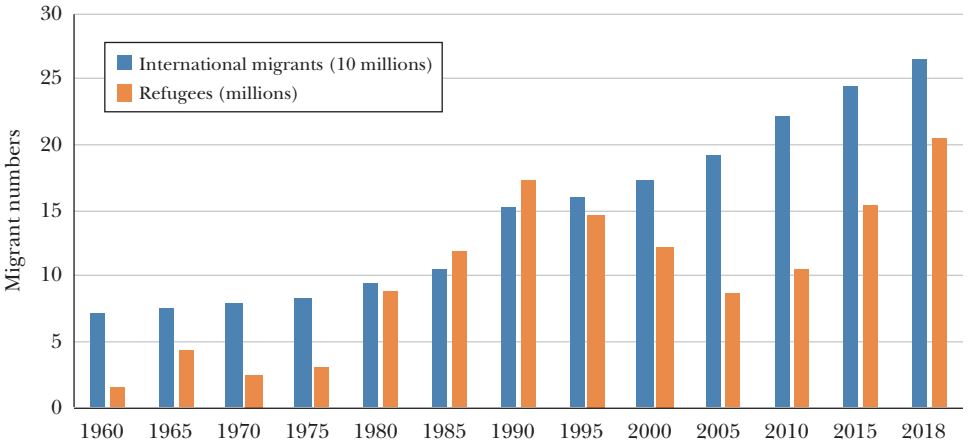
Most of those fleeing civil wars and human rights abuses are forcibly displaced within their own country or seek refuge in a country nearby. But the focus here is on asylum seekers who have grabbed the headlines and created public debate by migrating to the stable, safe, and secure countries of the West. Asylum migration has a long history, but the number arriving at the doors of the rich world has been on the increase. In 2015–2016, more than a million migrants from Syria and other Middle Eastern and Asian countries sought entry to the European Union and, from 2018, migrant caravans traveling from Central American countries converged on the US border with Mexico. So what explains asylum migration, and how does it differ from other migrations? And how have policies towards refugees and asylum seekers evolved in response to changing social and political pressures?

In this paper, I begin by presenting long-term trends on the number and composition of refugees and asylum seekers. The following section examines the political and institutional history that has drawn an increasingly sharp distinction between refugees and other types of migrants. Recent analysis has explored the determinants of asylum migration and has attempted to evaluate the effects of policies such as tighter border controls and more restrictive evaluation of asylum applications. Against this background, I examine how changes in public opinion and politics are shaping asylum policies in the aftermath of recent surges in asylum applications.

How Many Refugees?

The United Nations High Commissioner for Refugees (UNHCR) estimates the total number of refugees worldwide at the end of 2018 at 20.1 million. This is less than one-third of the total of 70.8 million "forcibly displaced persons," which also includes those displaced within their home country (41.3 million) and Palestinians

Figure 1
Worldwide Migrants and Refugees since 1960

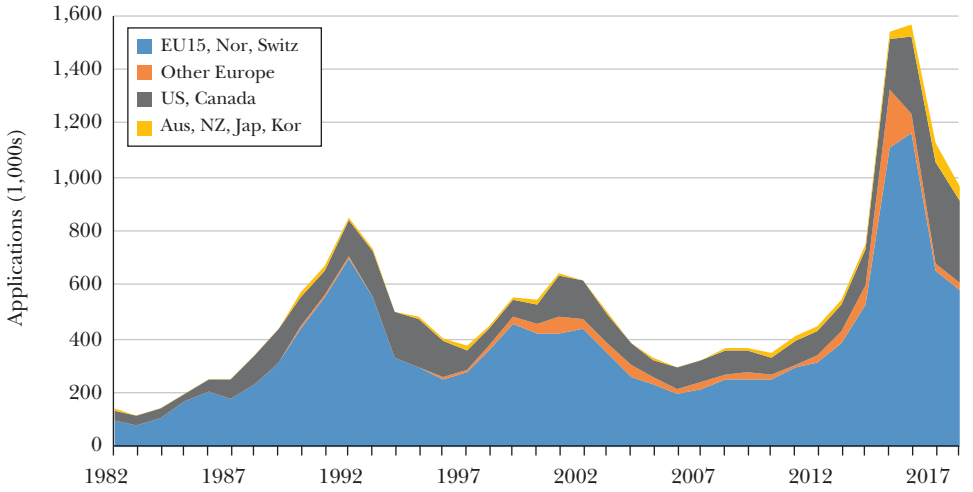


Source: Migrants: World Bank, International Migrant Stock. Refugees: 1960 to 1995 from UNHCR, State of the World's Refugees (2000) Annex 3; 2000 to 2010 from UNHCR Statistical Yearbook for 2007, Annex Table 20, and 2014, Annex Table 25; 2015 and 2018 from UNHCR Global Trends for 2015 and 2018, Annex Table 1.
Note: End year totals of international migrants (in 10 millions) and refugees (in millions).

(5.5 million) who come under a separate mandate (UNHCR 2019, 2). In 2018, refugees were 7.6 percent of the stock of all international migrants (defined as those living outside their country of birth). As Figure 1 shows, the number of refugees grew faster than total migrants from 1960 to 1990. After declining to the mid-2000s, the total number of refugees has risen steeply, largely as a result of conflicts in Syria, South Sudan, and Myanmar. As of 2018, two-thirds of refugees are from just five countries: Syria, Afghanistan, South Sudan, Myanmar, and Somalia. Of the total, 85 percent of refugees are located in developing countries, often just across the border from the origin country, and about 30 percent of these languish in organized refugee camps.

Each year, a small proportion of those recently displaced arrive as asylum seekers at the door of high-income Western countries in the hope of gaining recognition as refugees. In 2018, they were just 7 percent of those newly displaced (most of whom were internally displaced). The vast majority of asylum applicants in the developed world arrived as "spontaneous asylum seekers," having migrated from the origin country on their own initiative and not as part of an organized program. In contrast, the number of refugees who were transferred directly from refugee camps through resettlement programs averaged less than 100,000 until recently, but increased to a temporary peak of 189,000 in 2016. In 2016, 51 percent of resettled refugees went to the United States and another 39 percent went to Australia and Canada, while Europe took less than 10 percent. Since the late 1980s, the overwhelming majority of spontaneous asylum seekers have arrived in Europe. A large proportion gained

Figure 2
Asylum Applications to Western Countries, 1982–2018



Source: 1982 to 2000 from UNHCR, Statistical Yearbook for 2001, tables C1 and C2; 2001 to 2013 from UNHCR, Asylum Levels and Trends, 2005, 2009, and 2013, table 1; 2014 to 2018 from OECD, International Migration Outlook 2019, table A3.
Note: Annual number of persons applying for asylum, excluding repeat applications and appeals.

unauthorized entry, often traversing continents and traveling by hazardous land and sea routes. Frontex, the European Union's combined border force, estimated that unauthorized border crossings into the European Union increased from 105,000 in 2009 to a peak of 1.82 million in 2015.

Figure 2 shows the annual number of new asylum claims lodged in Europe, North America, Australasia, and Japan/Korea over the last 37 years. Most of the long-run increase is accounted for by asylum applications to Europe, which received 76 percent of total applications over the 37-year period, and especially Western Europe (71 percent). The total numbered less than 200,000 until the mid-1980s then rose steeply to a peak in 1992. This was the result of a surge of applications from Asia in the aftermath of the Vietnam War, followed by an even larger increase in applications, mainly from and through Eastern Europe, that attended the fall of the Berlin Wall and the dissolution of the Soviet Union. It was followed a decade later by a wave of applicants fleeing the Kosovo conflict. But what stands out above all is the steep increase during the Syrian crisis to one and a half million applications per annum in 2015–2016.

Most asylum applicants come from low-income countries embroiled in civil wars, internecine strife, and human rights abuses. Table 1 shows the top 30 origin countries by total applications over the decade 2009–2018. The Middle East, Africa, and Asia are the most prominent source regions, but there are also important origin countries in Europe (Serbia, Russia, and Albania) and in Latin America (El Salvador, Mexico, Guatemala, and Venezuela). China and India appear on the list, even

Table 1
Asylum Applicants to Western Countries by Origin: Total, 2009–2018

Origin country	*Total (000s)*	*Origin country*	*Total (000s)*	*Origin country*	*Total (000s)*
Syria	1,098.9	Albania	183.7	Georgia	97.4
Afghanistan	629.7	El Salvador	180.7	Guinea	87.3
Iraq	429.0	Somalia	176.1	Sri Lanka	84.0
Serbia	295.4	Mexico	160.4	Ukraine	79.3
Pakistan	275.2	Guatemala	138.3	Dem. Rep. Congo	71.0
Nigeria	252.8	Venezuela	133.9	Gambia	71.0
Eritrea	244.9	Bangladesh	123.8	Algeria	70.4
China	244.4	Honduras	109.9	Haiti	70.2
Russia	212.2	Turkey	106.9	Sudan	65.2
Iran	201.1	India	99.9	Mali	64.2

Source: Calculated from OECD, International Migration Database.
Note: Asylum applications from the top 30 origin countries to the EU28 plus Australia, Canada, Japan, South Korea, New Zealand, and the United States over the decade 2009 to 2018.

Table 2
Asylum Applicants to Western Countries by Destination: Total, 2009–2018

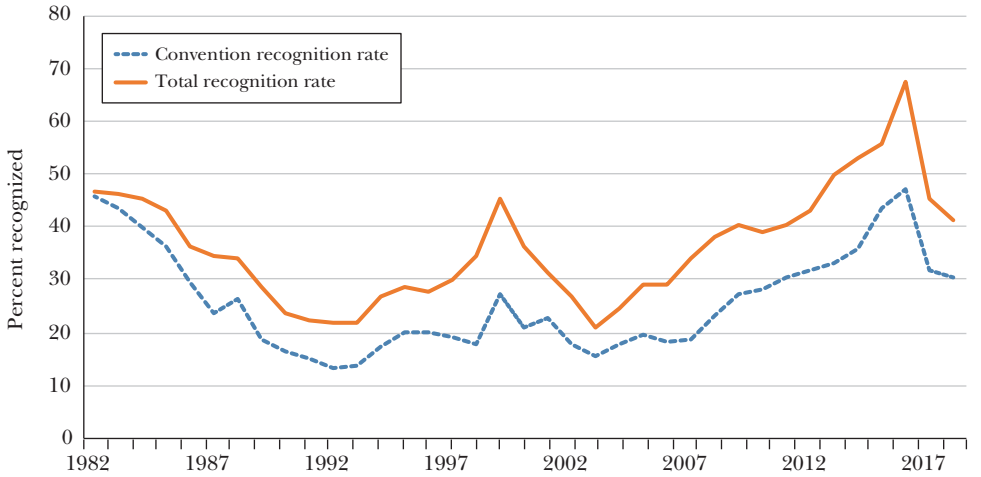
Destination country	*Total (000s)*	*Per 1,000 population*	*Destination country*	*Total (000s)*	*Per 1,000 population*
Germany	1,986.4	24.4	Switzerland	210.7	25.9
United States	1,462.1	4.6	Belgium	195.4	17.6
France	665.1	10.1	Netherlands	185.8	11.0
Italy	553.9	9.2	Australia	167.5	7.2
Sweden	478.1	49.3	Spain	135.0	2.9
United Kingdom	318.0	4.9	Norway	109.9	21.6
Hungary	276.6	28.0	Poland	77.4	2.0
Canada	270.5	7.7	Denmark	75.2	13.3
Austria	263.9	30.8	Finland	67.4	12.4
Greece	245.9	22.5	Japan	63.5	0.5

Source: Calculated from OECD, International Migration Database.
Note: Asylum applications over the decade 2009–2018 from all origin countries. The numbers in this table include applicants who were stateless or of unknown nationality. The figure for the United States includes both affirmative and defensive applications and has been adjusted by the OECD to reflect the number of individuals.

though the number of applications is small relative to their populations. Table 2 reports applications for the top 20 destination countries. Germany and the United States received the largest number of applications over the decade, but relative to population, the leading country is Sweden with 49 applicants per 1,000 population, followed at some distance by Austria, Hungary, and Switzerland.

Asylum applicants enter into a process to determine whether they qualify as refugees under the definition in the 1951 Refugee Convention or are eligible for admission on other humanitarian grounds. The total recognition rates (Convention

Figure 3
The Refugee Recognition Rate for 24 Countries, 1982–2018



Source: 1982 to 2005 from UNHCR, Statistical Yearbook for 2001 tables C26 and C29, and 2005 tables C27 and C30; 2006 to 2018 from UNHCR, Global Trends for 2006 to 2018, table 10.
Note: The countries included in the weighted recognition rates are: the EU-15 (excluding Luxembourg), the Czech Republic, Hungary, Poland, Norway, Switzerland, Australia, New Zealand, Japan, Canada, and the United States.

plus humanitarian) for 24 countries since 1982 are plotted in Figure 3. These recognition rates were close together in the early 1980s, but a gap emerges as weaker forms of recognition were adopted in the face of the rising number of applications. The average total recognition rate over the most recent 37 years is just 36 percent (26 percent for Convention recognition). Even if successful appeals were taken into account, the share of those applying for asylum who receive some form of recognition would not exceed half. Unsuccessful applicants are legally required to leave the country either voluntarily or by deportation, although a significant proportion disappear into the informal economy and remain as undocumented immigrants.

The total recognition rate peaked in 1999 and again in 2016. This pattern reflects both variations in the gravity of asylum claims and changing policy towards them. In Europe, the shift towards tougher policy in the early 2000s was arrested in the following years as the EU's Common European Asylum System came into effect. Against this background, the rising number of asylum claims recognized as valid increased, fueled by the so-called "Arab Spring," to reach a crescendo in the Syrian migration crisis of 2015–2016, when the existing policies were temporarily suspended. The sharp fall in asylum applications after 2016 largely reflects the agreement between the European Union and the main transit country, Turkey, which stemmed the flow across the Aegean Sea. The decline in recognition rates represents a return to preexisting policies. But both the volume of applications and the average recognition rate remain high by historical standards as the underlying

pressures persist. It remains to be seen whether this really is a "paradigm shift," as sometimes suggested (UNHCR Global Trends 2015, 3).

Evolution of the International Refugee System

For centuries, those facing oppression and persecution—often on religious grounds—have sought sanctuary in other countries. From the Huguenots in the seventeenth century to the Russian Jews in the late nineteenth century, these groups moved in modest numbers and generally with little hindrance (Marrus 1985, chap. 1). Within Europe and the New World, border controls were minimal and the authorities made no formal distinction between those fleeing persecution and other migrants. However, there were tight restrictions on migration from Asia and from colonial dependencies. After World War I, more restrictive and selective immigration policies were accompanied by the widespread introduction of passports as proof of identity. From that time, refugees emerged as a category distinct from other migrants. In the United States, immigration quotas by country of origin, introduced in 1921 and tightened in 1924, drastically restricted immigration from countries, some of which became sources of refugees. From then until 1952, refugees were neither formally included in immigration policy nor recognized separately.

From 1920 to 1950, the international refugee regime evolved through several stages (Hathaway 1984). Refugees were initially considered to be those who had been displaced by war and only later as those facing individual persecution. The initial focus was on providing legal status for stateless Europeans in response to mass displacements across shifting borders in the aftermath of World War I. These included two million Poles and a million Germans as well as many thousands of Magyars, Greeks, and Armenians. In 1921, the newly established League of Nations created a High Commissioner for Refugees with a mandate to assist, firstly, displaced Russians and then other nationalities by negotiating the exchange, repatriation, and resettlement of refugees, one key element of which was the issue of internationally recognized travel documents.[1] With the rise of Fascism, the focus shifted in the 1930s from the effects of displacement to the causes of persecution as group-specific mandates were issued, one of which applied to Jews fleeing Austria and Germany. The United States eased its eligibility criteria to admit a few thousand (but still did not fill its German quota until 1939), while more found sanctuary in France. But international diplomacy aiming to resettle larger numbers failed, and increasingly restrictive immigration policies around the world meant that there were few other havens for refugees (Loescher 2001, 31; Marrus 1985, chap. 3).

World War II created even greater displacement. By 1945 there were over 30 million displaced persons in Europe, not counting the 13 million ethnic Germans expelled mainly from Czechoslovakia, Poland, and the Soviet Union. At the end of the war, voluntary and official agencies assisted eight million European refugees, but a million more remained displaced. The initial focus on exiles from Fascism

[1] The first High Commissioner of Refugees, polar explorer Fridtjof Nansen, instigated the issuing of identity certificates, which became known as Nansen passports.

and Nazism then transformed into concern with those fleeing communism. The International Refugee Organization, created in 1946, was an initiative of the United States against Soviet opposition, and it specifically sought to distinguish between those fleeing persecution and those migrating for other reasons. It set out a definition of a refugee, which focused on the individual rather than the group and on the expectation of future persecution rather than on the circumstances of past displacement. It also reflected a shift from viewing repatriation as the principal solution to refugee problems to establishing a role for permanent resettlement elsewhere. The successor organization to the International Refugee Organization, the Office of the United Nations High Commissioner for Refugees (UNHCR) created in 1949, was followed in 1951 by the UN Convention Relating to the Status of Refugees. The Convention built upon the 1948 Universal Declaration of Human Rights, which included in Article 14 the right to seek asylum from persecution (Goodwin-Gill 2008), and following the precedent of the International Refugee Organization, it enshrined individual fear of persecution as the criterion.

The Refugee Convention (UNHCR 1951) includes three interlocking elements, which have shaped refugee policy up to the present. First, a signatory state must offer a procedure to assess whether or not each individual lodging a claim qualifies as a refugee under the Convention's definition of being outside that person's origin country and having a "well-founded fear of persecution" (Article 1). Second, while being on a country's territory (or at the border) does not, of itself, guarantee access to the process, the so-called *non-refoulement* clause (Article 33(1)) forbids returning a person to a place where that person's life or freedom would be threatened. Third, illegal entry or presence in the country does not prejudice admission to the procedure for determining refugee status or the outcome of that process (Article 31). In addition, while the Convention does not provide the right to permanent residence, it does encourage host countries to "facilitate the assimilation and naturalization of refugees" (Article 34). The Convention originally applied only to those displaced in Europe before 1951, but its scope was radically widened by the 1967 New York Protocol, which removed geographic and time limitations. It was gradually adopted worldwide and the number of signatory states increased from 60 in 1970 to 145 in 2015. It is noteworthy that, in principle, there is no limit to the number of asylum applications a state is obliged to process and accept.

The United States did not sign the 1951 Convention, and its policies diverged from those of Europe. Instead, it developed a series of initiatives, such as the 1952 Escapee Program, which focused on refugees from the Soviet Union and Eastern Europe. During the early years of the Cold War, refugees moving to the West were welcomed as a powerful symbol of Western superiority over communism, especially in the United States. Reflecting Cold War strategy, the bulk of refugees admitted to the United States during this period were from communist countries and were admitted for resettlement through executive orders outside of the immigration quota (Zucker and Zucker 1996, chap. 2). In the 1970s, the human rights agenda gained increasing popular support as, in the wake of the Vietnam War, the media fed public awareness of oppression and international conflicts in Latin America, Asia,

and Africa. This was reflected in growing support for humanitarian agencies such as Amnesty International, which won the Nobel Peace Prize in 1977, and Human Rights Watch, which was launched in the United States in 1978 (Neier 2012). It was also reflected in public policy: the Jackson-Vanik Amendment in 1974 (trade sanctions against nonmarket countries that denied the right to emigrate) and the creation of an Assistant Secretary of State for Human Rights in 1977. Finally, the US Refugee Act of 1980 established an annual refugee quota of 50,000, and in principle shifted the emphasis from country of origin to the plight of the individual, aligning more closely with the Refugee Convention.

The Refugee Act widened the scope of US refugee policy, and it provided a procedure for refugee status determination, which was foreshadowed by Canada in 1976 and Australia in 1978. In these countries, while the main mechanism was resettlement direct from countries of first asylum, the door was also opened to spontaneous asylum seekers. The United States nevertheless continued with ad hoc measures and a focus on exiles from communism; for example, Cubans were favored over Haitians and Nicaraguans over Salvadorans and Guatemalans (UNHCR 2000, 174–77). But of the two million that the United States resettled from 1975 to 1999, two-thirds were from Vietnam, Cambodia, and Laos. The Vietnamese boat people symbolized what was to follow, as the relatively liberal refugee regime of the 1980s faced severe challenges with growing numbers of spontaneous asylum seekers, often arriving illegally and from ever-more remote parts of the world. The end of the Cold War, heralded by the fall of the Berlin Wall and the dissolution of the Soviet Union, generated a surge in the numbers seeking asylum, just as the strategic value of refugees receded (Zucker and Zucker 1996, 37–38). In Europe, the steep increase up to 1992 (shown in Figure 2) led to tougher policies that included visa restrictions and tougher status determination policies (shown in Figure 3). Most notable was the 1992 amendment to Germany's Basic Law providing that asylum claims by applicants who originated from safe countries of origin or who traveled though safe third countries were deemed to be manifestly unfounded (Hailbronner 1994). Across the Atlantic, the US Illegal Immigration Reform and Immigrant Responsibility Act of 1996 restricted access to asylum procedures for those arriving without documents.

The further round of policy tightening that took place from the early 2000s in the face of rising applications was precipitated by the attacks of September 11, 2001. This intensified concerns that asylum seekers from conflict-ridden countries presented not only an economic burden and social problem but also a security risk. The USA Patriot Act of 2001 increased border security and identity checks, and sweeping reforms were also introduced in Australia (2001) and Canada (2002). In Europe, stricter border controls and visa policies were aimed at denying access while tougher processing policies and less generous welfare provisions were used to deter prospective applicants. But the first round of directives in the EU's Common European Asylum System in the mid-2000s sought to prevent a race to the bottom in asylum policies by harmonizing policies and striking a balance between excluding economic migrants while protecting the rights of genuine refugees. Even though asylum policies have become more restrictive

since the 1980s, there has been no mass defection from the Refugee Convention, a treaty that was conceived in conditions very different from today. Thus, the key elements of the liberal post-World War II regime—the right to claim asylum and the *non-refoulement* provision—remain in place without regard to the numbers that this may imply. The European migration crisis of 2015–2016 and the migrants gathering on the US southern border since 2017 have put these principles under severe pressure and have opened once again the question of whether existing asylum policies are still fit for current purposes.

What Drives Asylum Applications?

Existing studies have identified key factors that influence the number of refugees. Davenport, Moore, and Poe (2003) found that the stock of refugees around the world could be explained mainly by genocide, civil war, dissident conflicts, and political regime transitions. Consistent with this, worldwide refugee numbers run parallel with indicators of conflict, which ascend steeply to a peak in 1992 and then decline before reversing from 2011 (Center for Systemic Peace 2018). Recent examples include the war in Syria and persecution of the Rohingya in Myanmar, but while the first produced large outflow to the West, the latter did not. In a study of bilateral refugee movements, Moore and Shellman (2007) found that, while most migrants moved to contiguous countries, movements beyond countries of first asylum were positively related to the locations of previous migrants, but were constrained by the costs of migration. Annual asylum applications to the developed world have increased on trend relative to the worldwide refugee stock as more migrants have moved beyond countries of first asylum. Taking the ratio of asylum applicants to the developed world (shown in Figure 2) to the world refugee stock (shown in Figure 1) as 100 in 1985, this index increased to 345 in 1995, 272 in 2005, and 674 in 2015.

Several studies have assessed the push and pull forces behind asylum applications to industrialized countries by analyzing panel data on the number of applicants by origin, by destination, and over time. The most important origin-country variables are political terror and lack of civil liberties; civil war matters less, perhaps because war *per se* does not necessarily confer refugee status (Hatton 2009, 2017a). There is weaker evidence that declines in origin-country income per capita leads to more asylum applications, which offers modest support to the view that economic migration is part of the story. Proximity and access are important in determining the volume of asylum applications. Countries that are small but nearby can generate large flows—as with a quarter of a million Cubans moving to the United States in the 1970s and 400,000 Serbians and Montenegrins moving to the European Union in 1995–2004—provided that the door is left ajar. But the growth of transit routes and migrant networks have fueled the upward trend of applications from more distant origins. For example, travel in caravans through Mexico combined with violence and drought at home, a growing diaspora, and mixed messages about future US policy all combined to boost migration from Central America (Capps et al. 2019).

How does asylum migration differ from migration though other channels? Studies of total migration flows—including both asylum and non-asylum migration—that share the same panel data structure produce similar but not identical findings. The most obvious difference is the much greater influence on asylum migration of terror and human rights abuse in origin countries. Another difference is that economic "pull factors" in destination countries are stronger and "push factors" from origin countries are weaker for non-asylum migration. For example, Mayda (2010) and Ortega and Peri (2013) report large and significantly positive effects of destination-country income per capita on migration, but smaller and sometimes insignificant negative effects of origin-country income.[2] As in many migration models, the most powerful single variable influencing asylum-seeker flows to a country is the stock of previous migrants from the same origin. Underlying these network effects are historic factors shaping migration such as colonial ties, common language, and shared culture, as well as geographic proximity. The negative effect of distance is especially important for asylum applications, and it matters even in the presence of the migrant stock, something that probably reflects the greater costs and hazards of what, for many, is risky clandestine migration.

A particularly important issue is whether, and to what extent, restrictive asylum policies reduce asylum applications, especially as these are often purposely designed for deterrence. Policies that may influence the volume of asylum applications can be divided into three types. First, policies such as border surveillance, visa policies, and carrier sanctions seek to deny admission to asylum procedures by restricting access to the border. In the European migration crisis of 2015–2016, countries in the EU's eastern border adopted strict controls on border crossing and admission to asylum procedures. Second, rules that are applied in processing asylum claims can influence the likelihood that an applicant gains recognition. For example, when in 2013 Sweden granted all Syrian asylum seekers permanent instead of temporary residence, the number of applications more than doubled (Andersson and Jutvik 2019). Third, restrictions on movement that apply during processing and cuts in welfare benefits, such as the 47 percent benefit cut introduced by Denmark in 2015, might also deter asylum applications.

These polices are hard to quantify, but they can be crudely represented by an index comprising dummy variables for changes in each subcomponent of policy. When these variables are included in a model of asylum applications, border controls and processing policies have significant deterrent effects while welfare policies do not (Hatton 2004, 2009, 2017a). One interpretation is that what matters most to asylum seekers is the prospect of gaining permanent settlement, whatever the short-term hardships. The wave of tougher border controls and processing

[2]A related issue is that most migrants are young, which is predicted by economic theory because the net present value of investing in migration is greater the longer the duration of expected future returns. Consistent with this insight, studies of migration find that emigration is greater the larger are the young cohorts (aged 15–29) in the origin country (Mayda 2010; Hatton and Williamson 2011; Hanson and McIntosh 2016). That may also be true of asylum seekers, most of whom are young, but this effect has not been thoroughly investigated.

policies that took place between 1997 and 2005 reduced applications to 19 major destination countries by nearly 30 percent. From 2005 to 2014, countries such as the United States, Canada, and the United Kingdom continued to tighten their policies while others, including France, Italy, and especially Sweden, eased theirs, so that the overall effect was a modest increase in applications (Hatton 2017a, 464). In this light, it is not surprising that subsequent dramatic policy shifts in Europe had sizable effects on the volume of applications. The diverse incentives and deterrents can influence the characteristics of asylum applicants as well as the overall number. A study of migrants crossing the central and eastern Mediterranean routes in 2015 and 2016 found that those who claimed to be fleeing persecution were more positively selected on education than economic migrants (Aksoy and Poutvaara 2019). Those with low education were more often heading for countries with easier access to employment and more generous welfare states, but such intentions were also influenced by rising border restrictions on different routes.

It is sometimes suggested that more restrictive policy adopted by one country simply deflects asylum applicants to others. For regular migration, there is some support for this view (Ortega and Peri 2013), and this might be particularly important in the European Union, where nearby countries could be close substitutes. A careful test supports the deflection effect on asylum applications to third countries but finds it to be small (Barthel and Neumayer 2015). A possible reason is that the EU's so-called Dublin Regulation (which was suspended in 2015–2016) requires that an applicant can lodge an asylum claim in only one country, normally the country of first arrival, which restricts potential access to asylum procedures at alternative destinations. It has also been suggested that more restrictive policies on other types of immigration could *increase* asylum flows to a country, as potential immigrants seek an alternative immigration channel. Here, too, the evidence supports a substitution effect. However, employment-based immigration policies became *less* restrictive on average in 19 major destinations from 1997 to 2014, and this reduced asylum applications on average by 9 percent (Hatton 2017a, 463).

The effects of border controls are likely to be heterogeneous. Much of the evidence comes from the experience on the US-Mexico border. In the 1980s and 1990s, undocumented migration across this border increased in tandem with manpower and expenditure on border control, suggesting that policy had little effect. Increasing apprehension rates at the main crossing points diverted migrants to other sectors where access is more difficult, which raised the cost of employing smugglers ("coyotes") but had only modest effects on the total number of attempted crossings (Gathmann 2008; Massey, Durand, and Pren 2016; Lessem 2018). Indeed, the majority of those apprehended were granted voluntary return to Mexico, only to repeat the attempt until successful, while those who crossed successfully were less likely to return. But barriers were strengthened and surveillance intensified further, and from 2005 on, tougher sanctions were imposed that included criminal proceedings. Analysis of individual-level data on apprehensions for 2008–2012 indicates that this reduced the probability of re-apprehension within a year by nearly one-quarter (Bazzi et al. 2018). With the subsequent transition from single Mexicans looking

for work to Central American families seeking asylum, the United States faces new challenges at the border.

Unauthorized crossings to Europe have long been made with the intention of applying for asylum and gaining permanent residence. Most of these migrants are from countries that do not share a land border, so that unauthorized travel often involves the costs and risks of long and difficult migration routes through other countries and/or across the Mediterranean. The changing importance of different migration routes to the European Union during the last decade is largely a result of the vagaries of enforcement policies at different crossing points, rather than of substitution between routes by migrants (Hatton 2017a, 475–79). A good example is when the "friendship agreement" of 2008 between Italy and Libya collapsed with the demise of the Gaddafi regime in 2011. This increased unauthorized migration through the central Mediterranean route between 2010 and 2012 by a factor of three. Friebel et al. (2018) show that the increase in actual and intended migration came from countries relatively near Libya. There was almost no increase in migration, actual or intended, from more distant countries such as those in the Middle East and no reduction in travel through other routes.

Perhaps the most dramatic recent example of enforcement effects is how the massive surge of migrants through the western Balkans and eastern Mediterranean, as a result of the war in Syria, was brought to an abrupt halt after the 2016 agreement between the European Union and Turkey. The number of unauthorized crossings through the western Balkans and eastern Mediterranean fell from 1.65 million in 2015 to 54,500 in 2017, with only modest effects on the numbers traveling through other routes. Although the number crossing from Libya to Lampedusa (Italy) and Malta remained high, most of these migrants were from sub-Saharan Africa and the three leading nationalities were Nigeria, Guinea, and Côte d'Ivoire (Frontex 2018, 43). This experience indicates that land and sea crossings can be stemmed, but only with draconian policies and in cooperation with transit countries.

Public Opinion, Politics, and Policy

The dramatic increase in asylum applications in recent years has created headlines and alarmed policymakers. There is a widespread perception that public opinion has shifted dramatically against immigrants in general and asylum seekers in particular. This has been linked with increasing support for populist political parties, particularly those of the far right. Even when such parties do not get into government, they may shift the agendas of mainstream political parties towards a more anti-immigration stance.

What does survey evidence show on how public opinion has shifted? In 2002, 2014, and 2016, the European Social Survey (ESS) asked respondents if they agreed/disagreed with the statement: "the government should be generous in judging applications for refugee status." The first row of Table 3 reports the average over 17 countries of the proportion of respondents that disagreed or strongly

Table 3
Anti-refugee and Anti-immigration Opinion in 17 European Countries

	2002	*2014*	*2016*	*Change 2002–2014*	*Change 2014–2016*
Applicants for refugee status (% disagree or disagree strongly)	40.9	26.6	36.1	–14.3	9.5
Immigrants of different race/ethnic group (% few or none)	48.3	42.3	41.8	–5.9	–0.5
Immigrants from poor countries (% few or none)	47.8	50.4	43.9	2.6	–6.5

Source: European Social Survey, cumulative file.
Note: The first row is the percentage of respondents who "disagreed" or "disagreed strongly" with the statement: "the government should be generous in judging applications for refugee status." The second and third rows are the percentages of respondents who replied "a few" or "none" to the question: "to what extent do you think [country] should allow. . . people of a different race or ethnic group from most [country] people" and ". . . people from the poorer countries outside Europe." These are the unweighted averages for the following countries: Austria, Belgium, Czech Republic, Finland, France, Germany, Hungary, Ireland, the Netherlands, Norway, Poland, Portugal, Slovenia, Spain, Sweden, Switzerland, and the UK.

disagreed with the statement. From 2002 to 2014, on average, there was a fall in the proportion of those expressing anti-refugee sentiment by 14.3 percentage points. In 2014, anti-refugee preference averaged 26.6 percent, and it was less than 50 percent in all 17 countries, ranging from 7.6 percent in Portugal to 47.0 percent in the Netherlands. But from 2014 to 2016, the decline in anti-refugee sentiment was sharply reversed everywhere except Ireland, Spain, and the United Kingdom. In Germany, anti-refugee sentiment increased 17 percentage points and in Hungary by 26 percentage points. Trends in opinion on immigration policy are rather different, even towards otherwise similar groups such as immigrants from minority ethnic backgrounds and those from poorer countries outside Europe. Anti-immigration responses are taken as the percentage who prefer admitting "a few" or "none," compared with the alternatives "many" or "some." As Table 3 shows, from 2002 to 2014, there was much less decline in negative sentiment towards immigrants as compared with refugees. There was some softening of views towards ethnic minority immigrants but not towards those from poor countries, with some reversal of trends from 2014 to 2016.

The United States presents a somewhat different picture. Each June, Gallup asks if immigration should be kept at its present level, increased, or decreased (Gallup 2014). The percentage of respondents wanting immigration to be decreased fell from 49 in 2002 to 41 in 2014, 38 in 2016, and 35 in 2019. Despite the growing support for immigration, there is evidence of increasing concern about the situation on the border with Mexico, which 74 percent of respondents in 2019 considered to be a "crisis" or a "major problem." But when asked about admitting refugees who have left Honduras and other Central American countries, 57 percent approved

while 60 percent either opposed or strongly opposed expanding the construction of walls along the US-Mexico border. In this respect, opinion in the United States has some parallels with that in Europe on the eve of the migration crisis.

Two important elements contribute to the overall climate of opinion towards asylum seekers. First, public opinion is very strongly against unauthorized entry. Among respondents to a survey of eight European countries in 2013, an average of 75 percent were "worried about illegal immigration," as compared with 29 percent who were "worried about legal immigration." For the United States, these figures were 61 percent and 25 percent, respectively.[3] It is likely that the increase in unauthorized arrivals has further hardened attitudes towards spontaneous asylum seekers. Second, and related to this, the *salience* of immigration has increased. Salience refers to how important a respondent thinks an issue is, as distinct from the respondent's position or preference over the issue (as reported in Table 3). One measure of salience is recorded in the Eurobarometer surveys, which ask respondents about the two most important issues facing the country. From 2004 to 2012, roughly 10 percent of those in the survey ranked immigration in their top two issues. But in 2015, this shot up to over 30 percent for the European Union as a whole and a whopping 75 percent in Germany.

Populist parties have been gaining influence across Europe, and although they vary widely in other ways, they typically share a strong anti-immigration stance. In Italy, votes for the centre-right coalition in the national elections of 2001–2008 were positively influenced by the proportion of foreign-born in the local population (Barone et al. 2016). In Austria, votes for the far-right Freedom Party in elections from 1979 to 2013 are causally related to the increase in immigration (Halla, Wagner, and Zweimüller 2017). In districts of Hamburg, Germany, voting for the far-right parties in state and national elections in 1987–2000 is linked to the share of immigrants (Otto and Steinhardt 2014). Across Europe, votes for nationalist parties in European elections are positively affected by the local share of low-skilled immigrants, especially those from outside Europe (Moriconi et al. 2018). These findings reflect both economic interests and cultural concerns, and they suggest that the (pro-immigrant) "contact effect" is overwhelmed by a "group threat effect," which reflects both fear of competition and cultural concerns. But these findings relate to immigration generally and not specifically to refugees or asylum seekers.

By exploiting the (exogenous) placing of refugees in localities in Denmark in 1986–1998, Dustmann, Vasiljeva, and Damm (2019) find causal evidence of a link between the presence of refugees and voting for anti-immigration parties in rural areas but the opposite effect in the main urban areas (consistent with group threat and contact effects, respectively). The recent refugee crisis of 2015–2016 fueled support for anti-immigrant parties, but this effect varied between countries and localities. In Upper Austria, support for the Freedom Party increased by less in municipalities that

[3]These figures were derived from the database for Transatlantic Trends 2013 (Stelzenmueller et al. 2013). In a 2014 Gallup poll, 77 percent of US respondents thought that controlling US borders to halt the flow of illegal immigrants into the United States was either "very important" or "extremely important" for government policy.

hosted refugee centers, but by more in border municipalities that migrants passed through on their way to Germany (Steinmayr 2018). Exposure to migrant arrivals on Greek islands also increased opinion in favor of exclusion and added electoral support for the far-right party, Golden Dawn (Hangartner et al. 2019; Vasilakis 2018; Dinas et al. 2019). This evidence suggests that, against a background of rising countrywide salience, contact or proximity to refugees mitigated or had mixed effects on the rise in voting for anti-immigrant parties, while direct experience of unauthorized migration boosted it.

There is much less evidence exploring the last link in the chain running from immigration to public attitudes and then on to changes in immigration policy. One strand of evidence suggests that higher public salience of immigration is associated with more restrictive asylum policies (Hatton 2017b). But because the legislative process is often protracted and the outcome uncertain, the immediate effects of shifting attitudes are more likely to be on enforcement within the existing policy framework. For example, surges in asylum applications are associated with slightly lower asylum-seeker recognition rates in European countries, but there is no clear relationship with the strength of far-right political parties in government (Neumayer 2005; Toshkov 2014).

In the European Union, the Common European Asylum System has increasingly constrained the policies of individual governments. But the migration crisis of 2015–2016, along with the collapse of border controls in southern Europe and Germany's short-lived open door policy pitched this policy regime into disarray. The public backlash against asylum migrants largely reflected concerns about unauthorized immigration, but it also presented an opportunity for further reform (Trauner 2016). The EU agreement with Turkey over the movement of Syrians, noted earlier, was followed in 2016 by the transformation of the EU's border force, Frontex, into a more integrated European Border and Coastguard Agency, with increased executive power and greater financial resources. The reforms also include a doubling of the EU's Asylum, Migration, and Integration Fund and the transformation of the European Asylum Support Office into a full-fledged EU Agency for Asylum, with greater operational powers. The crisis also led to measures to redistribute 170,000 asylum seekers from Greece and Italy, even in the face of opposition by four member states. This was a modest breakthrough for a policy of European burden-sharing that has long been discussed, but not acted upon.

Recent experience has led some to criticize as inefficient an asylum system that provides incentives to engage in risky unauthorized migration, only for the majority of such migrants to fail to gain recognition as refugees (Hatton 2017a). Tighter border controls reduce unpopular unauthorized migration, but they exclude both economic migrants and genuine refugees. An alternative would be more like the Australian system where tough border controls are accompanied by a resettlement scheme which, if scaled up on a per capita basis to the EU population, would admit around 375,000 refugees per year. Substituting resettlement for spontaneous asylum-migration was at the core of the EU-Turkey agreement, which provided that for every Syrian migrant returned to Turkey from the Greek islands, another Syrian refugee

would be resettled from Turkey to an EU member state. With that provision as background, in 2017, the EU adopted an expanded resettlement program of 50,000—or five times the number of the program launched in 2008. In contrast, the United States has moved in the opposite direction by reducing the resettlement target as the specter of spontaneous asylum-seeking increased. The US resettlement program of 96,900 in 2016, which was more than half of the worldwide total among developed countries, was reduced to just 22,900 in 2018.

Conclusion

Concern over refugees has increased in recent years as the numbers have surged. While most refugees are located in low-income neighboring countries where they first found asylum, the increasing number applying for asylum in the Western world has attracted widespread attention. These trends should be understood against the background of the evolution of international policy towards refugees and the changing incentives for asylum migration. The terms of the 1951 Refugee Convention and the asylum policies built upon it have provided clear incentives for spontaneous migration from poor, strife-prone countries to the developed world. While the evolution of policy sharpened the distinction between refugees and other immigrants, that difference has become increasingly blurred among asylum migrants.

Since the early 2000s, public attitudes towards genuine refugees have become more favorable, but concerns about unauthorized arrivals have increased. In Europe, these concerns came to a head in the migration crisis of 2015–2016, and the backlash from that experience has led to a range of policy reforms, particularly tougher border controls. But it also marked a small step towards favoring resettlement over spontaneous asylum-seeking. Meanwhile the United States has shifted the other way: with a leaky southern border and public support for the Central American refugees, the government has drastically cut its resettlement program. It remains to be seen whether the tougher border controls that have been proposed will in time be accompanied by a return to a more generous resettlement quota.

References

Aksoy, Cevat Giray, and Panu Poutvaara. 2019. "Refugees' Self-Selection into Europe: Who Migrates Where?" ifo Institute Working Paper 289. https://www.ifo.de/DocDL/wp-2019-289-aksoy-poutvaara-refugees-self-selection.pdf.

Andersson, Henrik, and Kristoffer Jutvik. 2019. "Do Asylum Seekers Respond to Policy Changes? Evidence from the Swedish-Syrian Case." https://www.dropbox.com/s/bzopi36wjhrs06r/Henrik4.pdf?dl=0.

Barone, Guglielmo, Alessio D'Ignazio, Guido de Blasio, and Paolo Naticchioni. 2016. "Mr. Rossi, Mr. Hu and Politics: The Role of Immigration in Shaping Natives' Voting Behavior." *Journal of Public Economics* 136 (4): 1–13.

Barthel, Fabian, and Eric Neumayer. 2015. "Spatial Dependence in Asylum Migration." *Journal of Ethnic and Migration Studies* 41 (7): 1131–51.

Bazzi, Samuel, Sarah Burns, Gordon Hanson, Bryan Roberts, and John Whitley. 2018. "Deterring Illegal Entry: Migrant Sanctions and Recidivism in Border Apprehensions." NBER Working Paper 25100.

Capps, Randy, Doris Meissner, Ariel G. Ruiz Soto, Jessica Bolter, and Sarah Pierce. 2019. *From Control to Crisis: Changing Trends and Policies Reshaping U.S.-Mexico Border Enforcement.* Washington, DC: Migration Policy Institute.

Center for Systemic Peace. 2018. "Global Conflict Trends: Assessing the Qualities of Systemic Peace." http://www.systemicpeace.org/conflicttrends.html.

Davenport, Christina A., Will H. Moore, and Steven C. Poe. 2003. "Sometimes You Just Have to Leave: Domestic Threats and Forced Migration, 1964–1989." *International Interactions* 29 (1): 27–55.

Dinas, Elias, Konstantinos Matakos, Dimitrios Xefteris, and Dominik Hangartner. 2019. "Waking Up the Golden Dawn: Does Exposure to the Refugee Crisis Increase Support for Extreme-Right Parties?" *Political Analysis* 27 (2): 244–54.

Dustmann, Christian, Kristine Vasiljeva, and Anna Piil Damm. 2019. "Refugee Migration and Electoral Outcomes." *Review of Economic Studies* 86 (5): 2035–91.

European Social Survey. "Cumulative File, ESS 1–8." Norwegian Centre for Research Data. https://www.europeansocialsurvey.org/data/ (accessed March 25, 2019).

Friebel, Guido, Miriam Manchin, Mariapia Mendola, and Giovanni Prarolo. 2018. "International Migration Intentions and Illegal Costs: Evidence Using Africa-to-Europe Smuggling Routes." CEPR Discussion Paper 13326.

Frontex. 2018. *Risk Analysis for 2018.* Warsaw, Poland: Frontex European Border and Coast Guard Agency. https://frontex.europa.eu/publications/?c=risk-analysis.

Gallup. "Immigration." https://news.gallup.com/poll/1660/immigration.aspx (accessed November 24, 2019).

Gathmann, Christina. 2008. "Effects of Enforcement on Illegal Markets: Evidence from Migrant Smuggling along the Southwestern Border." *Journal of Public Economics* 92 (10): 1926–41.

Goodwin-Gill, Guy S. 2008. "Convention Relating to the Status of Refugees, 1951, and the Protocol Relating to the Status of Refugees, 1967." UN Audio Visual Library of International Law. http://legal.un.org/avl/ha/refugees.html.

Halla, Martin, Alexander F. Wagner, and Josef Zweimüller. 2017. "Immigration and Voting for the Far Right." *Journal of the European Economic Association* 15 (6): 1341–85.

Hangartner, Dominik, Elias Dinas, Moritz Marbach, Konstantinos Matakos, and Dimitrios Xefteris. 2019. "Does Exposure to the Refugee Crisis Make Natives More Hostile?" *American Political Science Review* 113 (2): 442–55.

Hanson, Gordon, and Craig McIntosh. 2016. "Is the Mediterranean the New Rio Grande? US and EU Immigration Pressures in the Long Run." *Journal of Economic Perspectives* 30 (4): 57–82.

Hathaway, James C. 1984. "The Evolution of Refugee Status in International Law: 1920–1950." *International and Comparative Law Quarterly* 33 (2): 348–80.

Hatton, Timothy J. 2004. "Seeking Asylum in Europe." *Economic Policy* 19 (38): 5–62.

Hatton, Timothy J. 2009. "The Rise and Fall of Asylum: What Happened and Why?" *Economic Journal* 119 (535): 183–213.

Hatton, Timothy J. 2017a. "Refugees and Asylum Seekers, the Crisis in Europe and the Future of Policy." *Economic Policy* 32 (91): 447–96.

Hatton, Timothy J. 2017b. "Public Opinion on Immigration in Europe: Preference versus Salience." CEPR Discussion Paper 12084.

Hatton, Timothy J., and Jeffrey G. Williamson. 2011. "Are Third World Emigration Forces Abating?" *World Development* 39 (1): 20–32.

Hailbronner, Kay. 1994. "Asylum Law Reform in the German Constitution." *American University International Law Review* 9 (4): 159–79.

Lessem, Rebecca. 2018. "Mexico-U.S. Immigration: Effects of Wages and Border Enforcement." *Review of Economic Studies* 85 (4): 2353–88.

Loescher, Gil. 2001. *The UNHCR and World Politics: A Perilous Path.* Oxford: Oxford University Press.

Marrus, Michael R. 1985. *The Unwanted: European Refugees in the Twentieth Century.* New York: Oxford

University Press.

Massey, Douglas S., Jorge Durand, and Karen A. Pren. 2016. "Why Border Enforcement Backfired." *American Journal of Sociology* 121 (5): 1557–1600.

Mayda, Anna Maria. 2010. "International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows." *Journal of Population Economics* 23 (4): 1249–74.

Moore, Will H., and Stephen M. Shellman. 2007. "Whither Will They Go? A Global Study of Refugees' Destinations, 1965–1995." *International Studies Quarterly* 51 (4): 811–34.

Moriconi, Simone, Giovanni Peri, and Riccardo Turati. 2018. "Skill of the Immigrants and Vote of the Natives: Immigration and Nationalism in European Elections 2007–2016." NBER Working Paper 25077.

Neier, Aryeh. 2012. *The International Human Rights Movement: A History*. Princeton, NJ: Princeton University Press.

Neumayer, Eric. 2005. "Asylum Recognition Rates in Western Europe: Their Determinants, Variation, and Lack of Convergence." *Journal of Conflict Resolution* 49 (1): 43–66.

OECD. "International Migration Database." https://stats.oecd.org/ (accessed February 4, 2019).

OECD. 2019. *International Migration Outlook 2019*. Paris: OECD.

Ortega, Francesc, and Giovanni Peri. 2013. "The Effect of Income and Immigration Policies on International Migration." *Migration Studies* 1 (1): 47–74.

Otto, Alkis Henri, and Max Friedrich Steinhardt. 2014. "Immigration and Election Outcomes: Evidence from City Districts in Hamburg." *Regional Science and Urban Economics* 45 (C): 67–79.

Steinmayr, Andreas. 2018. "Contact Matters: Exposure to Refugees and Voting for the Far-Right." Unpublished.

Stelzenmueller, Constanze, Richard Eichenberg, Craig Kennedy, and Pierangelo Isernia. 2013. *Transatlantic Trends Survey*. Ann Arbor: Inter-university Consortium for Political and Social Research.

Toshkov, Dimiter D. 2014. "The Dynamic Relationship between Asylum Applications and Recognition Rates in Europe (1987–2010)." *European Union Politics* 15 (2): 192–214.

Trauner, Florian. 2016. "Asylum Policy: The EU's 'Crises' and the Looming Policy Regime Failure." *Journal of European Integration* 38 (3): 311–25.

United Nations High Commissioner for Refugees (UNHCR). 1951. *Convention and Protocol: Relating to the Status of Refugees*. Geneva: UNHCR.

UNHCR. 2000. *The State of the World's Refugees 2000: Fifty Years of Humanitarian Action*. Geneva: UNHCR.

UNHCR. *Asylum Levels and Trends in Industrialized Countries* (various years). Geneva: UNHCR.

UNHCR. *Global Trends: Forced Displacement* (various years). Geneva: UNHCR.

UNHCR. *Statistical Yearbook* (various years). Geneva: UNHCR.

Vasilakis, Chrysovalantis. 2018. "Massive Migration and Elections: Evidence from the Refugee Crisis in Greece." *International Migration* 56 (3): 28–43.

World Bank. *International Migrant Stock, Total*. https://data.worldbank.org/indicator/SM.POP.TOTL (accessed February 10, 2019).

Zucker, Norman L., and Naomi F. Zucker. 1996. *Desperate Crossings: Seeking Refuge in America*. Armonk, NY: M.E. Sharpe.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 94–121

The Labor Market Integration of Refugee Migrants in High-Income Countries

Courtney Brell, Christian Dustmann, and Ian Preston

Economic models of the integration of immigrants into a host society generally focus on two main categories of factors: what determines who chooses to migrate; and what determines the accumulation of human, social, and cultural capital after immigration. Along both dimensions, refugee integration is likely to differ considerably from that of the typical economic migrant (for discussion, see, for example, Becker and Ferrara 2019; Chin and Cortes 2015). In addition, the refugee experience itself adds complexity to the integration of these migrants, who have often experienced traumatic episodes in their country of origin or extended periods traveling or in temporary living situations (such as refugee camps) before arriving in the host country.

While economic migrants decide to relocate to another country based on the relative opportunities afforded abroad compared to at home, refugee migration—being forced and often unexpected—is driven by different factors, such as vulnerability to persecution and access to the wherewithal to enable flight. Refugees are therefore not economically selected to the same degree as economic migrants and have more limited ability to choose a specific destination to which they will migrate. As a result, refugees typically arrive in a host country with less locally applicable human capital, including language and job skills, than economic migrants and consequently are likely to start at significantly lower levels of wages and employability.

■ *Courtney Brell is a Researcher, Christian Dustmann is Director, and Ian Preston is Deputy Research Director at the Centre for Research and Analysis of Migration, University College London, United Kingdom. Dustmann and Preston are both Professors of Economics at University College London. Preston is a Research Fellow, Institute for Fiscal Studies, also in London, United Kingdom. Their email addresses are courtney.brell.17@ucl.ac.uk, c.dustmann@ucl.ac.uk, and i.preston@ucl.ac.uk.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.94.

After arrival, incentives for refugees to improve their economic prospects in the host country are mixed compared to economic migrants. On the one hand, beginning at a lower level of human capital means that the potential costs of investment (such as forgone wages) are lower, and the rate of return on this investment may possibly be higher (at least according to some views of how immigrants accumulate human capital). If these effects dominate, then refugees would be expected to undergo rapid assimilation, particularly early on in their stay. On the other hand, refugees often face an uncertain future. They do not know at first whether asylum will be granted, and even if it is, permission to stay may be explicitly temporary and subject to periodic reassessment with the possibility of revocation. Some refugees may wish to return to their home country as soon as it becomes safe to do so, but when that will become possible, if ever, is uncertain. Such uncertainty may reduce the incentives to invest in host-country-specific human capital, such as language or social networks, and this may inhibit the integration of this group (Adda, Dustmann, and Gorlach 2019). The uncertainty itself may also be psychologically distracting and a hindrance to integration.

Finally, the unique experiences of refugees will also affect their ability to integrate. Having experienced or witnessed conflict and persecution means that health issues, and particularly mental health issues, are common among the refugee population. The journey from their home to the host country, as well as potentially having been traumatic, may also have been long or involved extended stays in intermediate locations such as refugee camps. During this time, refugees' human capital may have deteriorated as they may have had few opportunities to perform productive work.

Taken together, these factors mean that the integration of refugees is likely to raise significant challenges. In this paper, we provide an overview of what is currently known about the economic integration of refugees into high-income host countries, and in particular into their labor markets. We begin with a discussion of some facts about the refugee experience prior to arrival in the host country—their flight, journey, and stays in intermediate locations.

Following this, we provide an overview of the labor market outcomes of refugees in a variety of developed countries, based on an unusually broad collection of existing micro data sources, supplemented by evidence from data made available to us by a number of authors who have studied the topic. We will illustrate significant heterogeneity in outcomes of refugees across different host countries, with the general pattern that refugees start off behind other immigrants in employment and wages, and while they catch up over time, this catch-up is more pronounced in employment rates than in wages. We also offer a nonexhaustive but illustrative overview of some of the recent research in this area.

Although our focus is on economic integration, and in particular labor market outcomes such as employment and wages, integration of immigrants into a society—whether refugees or economic migrants—ultimately has to do with a broad development of capacities for successful participation in the host society, supporting a sense of social belonging in the destination

country.[1] Moreover, these wider dimensions of integration are often important determinants of economic outcomes. Thus, we will also delve into some broader social factors: health, language skills, and social networks. These factors present particular challenges for the integration of refugees, and as such, finding ways for policy to take these challenges into account may help in easing the integration of refugees into the workforce and society as a whole.

We conclude with a summary and a discussion of insights for public policy in receiving countries with regard to refugees. The prospects for successful integration depend not just on actions of the refugee or the immigrant but also on the openness and specific policy choices of the receiving community. Many recipient countries have put considerable effort and expense into measures targeted at supporting refugees' absorption into their societies and economies, but it is not always clear that the outcomes of these policies are in line with prior expectations or justifications.

The Refugee Experience

The diversity of migrant experience means that telling individual stories risks portraying their details as representative, when in fact the real-life variety is beyond what it is possible to present through anecdotes or case studies. With that warning in mind, such stories can still be valuably illustrative and highlight some of the unique circumstances that refugees face. Before discussing the refugee experience in general terms, we briefly describe five individual refugee journeys, each anonymized but adapted from a documented story:[2]

Example A: A student and waitress lived with her husband and children in a refugee camp near Damascus for several years after their home was destroyed in the Syrian civil war. As fighting between opposing forces neared, they paid to be trafficked by bus to the Turkish border, a dangerous journey that involved passing through areas under the control of several rival groups. After a short period staying in a camp in Turkey, they risked a perilously overcrowded boat journey to Greece and from there proceeded mostly on foot across the Balkans, often hopping between camps on the way. After being trafficked across the Hungarian border, they were able to take a train to Munich and finally claim asylum there. Their journey lasted about a month.

Example B: A Rohingya family and their business were persecuted by the army in a village in Myanmar. After their home was confiscated, they fled their village and tried to establish a life elsewhere in Myanmar. Their son moved to study in

[1] For example, Harder et al. (2018) develop measures of integration along six dimensions: psychological, economic, political, social, linguistic, and navigational. The influential conceptual framework of Ager and Strang (2008) identifies ten domains of integration within four areas of attainment.

[2] The stories are loosely based on original reports available at Adams and Vinograd (2015), Alcorn (2019), Watson (2019), García (2019), and *Refugee Action* (2017).

Yangon where he distributed political pamphlets, for which he was arrested and tortured but secured release through bribery. Fearing further recrimination, he fled to Thailand and on to Malaysia where he spent nine years working as an unauthorized immigrant before being recognized by the United Nations as a refugee. He took a boat journey from Indonesia to Australia, which resulted in him being held for 32 months in an immigration detention center. A decade later he works in construction and for community organizations in Melbourne, but still awaits permanent protection status, and has little contact with his family.

Example C: A child was born in a refugee camp of some 200,000 people in Kenya, to which her parents had fled from the civil war in Somalia. She lived there for her first eight years with her parents, siblings, and father's other wives. She received little education and facilities in the camp were rudimentary. Her family was eventually selected for resettlement and moved to Baltimore where they remained for seven years, before relocating to Buffalo to be closer to relatives and a larger Somali community. She is now studying for a PhD in education.

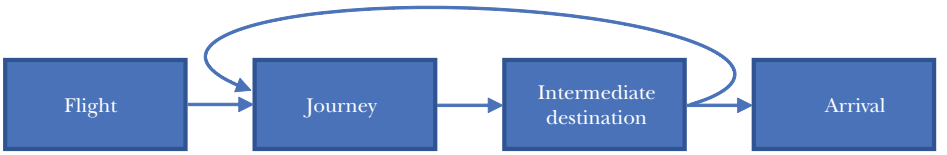
Example D: A mother of seven in a small community in Honduras participated in protests when water supplies to her village were compromised by a dam construction project. She was arrested and charged with trespassing, but the case was eventually dismissed. When a fellow protester was shot dead by police, she decided to leave with her two-year-old son and joined a migrant caravan traveling through Guatemala and Mexico to the US border, including a terrifying journey on top of a freight train. After crossing the border at Tijuana, she was held in detention for two weeks and spent a month in a shelter before relocation to Portland, Oregon, where she awaits a decision on her asylum application.

Example E: A young gay man moved to the United Kingdom from Algeria when his family discovered he was gay and tried to force him to marry his cousin. Struggling with depression, he stayed for several years with another cousin, overstaying his visa and helping with domestic chores while avoiding the formal economy. After learning from a charity that he might be eligible for refugee status, he applied for and was granted asylum. He now works as a sous-chef.

Of course, this small collection of individual stories encapsulates only a tiny proportion of the suffering and distress underlying refugee statistics. According to the UN High Commissioner for Refugees (2019a), in 2018, there were 70.8 million people forcibly displaced worldwide, including 25.9 million international refugees and 3.8 million individuals awaiting asylum decisions. For each one of these millions, there is an underlying story of hardship.

As the examples illustrate, the process of seeking refuge can have multiple stages, and at each stage, important decisions are made that will determine not only where and when a refugee will end up settling into a (semi-)permanent home, but also will influence their integration prospects after arrival. To structure our

Figure 1
The Stages of the Refugee Experience



discussion of these decisions and their potential consequences for refugee integration, we will break down the refugee path from origin to destination into the following stages as depicted in Figure 1: flight, journey, intermediate destination, and arrival.

Flight

During the past decade, the number of individuals displaced by war or persecution has increased dramatically, in large part due to ongoing conflicts in Asia and Africa (notably in Syria, Afghanistan, and South Sudan, which together have produced half of the global refugee and asylum-seeker stock in 2018; adding Myanmar and Somalia to this list accounts for two-thirds of global refugees) (UNHCR 2019a).[3] As the earlier examples illustrate, refugees may be fleeing civil conflict, religious or ethnic persecution, lethal police corruption, or inadequate protection of minority human rights.

The decision to flee one's home is traumatic, and even in the midst of ongoing conflict or persecution, many prefer to stay put. Aksoy and Poutvaara (2019) point out that, even if economic selectivity may be expected to be less strong for refugees than for other types of migrants, it will not be absent, and they show this using data for several countries. Wealth that would be abandoned in the home country upon flight will be a factor in the decision, as will economic prospects in possible destination countries. Of those that would like to leave, not all may have access to the resources needed to do so. In addition, persecution risk may be associated with economic prosperity (for example, if the persecution is motivated by perceived economic factors) and so may the risks associated with the journey (if the wealthier can buy their way out of dangerous situations or afford more reliable transport).

Nonetheless, if noneconomic factors have heightened importance for refugees, that may mean that refugee populations are likely to include both low- and

[3]We follow here the definition of a refugee from the UN High Commissioner for Refugees, which includes "individuals recognized under the 1951 Convention relating to the Status of Refugees, its 1967 Protocol, the 1969 Organization of African Unity (OAU) Convention Governing the Specific Aspects of Refugee Problems in Africa, the refugee definition contained in the 1984 Cartagena Declaration on refugees as incorporated into national laws, those recognized in accordance with the UNHCR Statute, individuals granted complementary forms of protection, and those enjoying temporary protection. The refugee population also includes people in refugee-like situations." In contrast, asylum seekers are "individuals who have sought international protection and whose claims for refugee status have not yet been determined…irrespective of when those claims may have been lodged" (UNHCR 2019a; for more detail, see Hatton, 2016, 2017, and forthcoming).

high-skilled individuals whose skills are more suited to their country of origin than to their destination country and demographic types who might be unlikely to migrate for economic reasons. This is not to say that refugees will not be distinctive in some respects since, as discussed, they will still be selected in other ways. Additionally, if there is heterogeneity in individual economic and cultural adaptability, then refugees (unlike economic migrants) will also not be selected in those terms, and this could tend to inhibit rapid integration.

Journey

Many of those displaced by conflict or persecution remain in their country of origin. In fact, of the stock of displaced persons recorded by the UN High Commissioner for Refugees (2019a) as of 2018, only 42 percent were refugees and asylum seekers; the remaining 58 percent being internally displaced. Many are displaced to nearby countries: nearly four-fifths of refugees live in countries neighboring their country of origin. These nearby destinations are typically developing; only 16 percent of refugees are hosted by countries in developed regions. Thus, as well as the decision to flee, refugees arriving in developed countries are often selected by having undertaken an especially long and difficult journey in search of a better life.

The details of a refugee's journey may differ hugely, and many choices are made along the way. Some paths are well understood by those taking them to have significant risks of death: for example, the UN High Commissioner for Refugees (2019b) reports that in 2018, with 141,000 Mediterranean arrivals to Europe, there were nearly 2,300 estimated dead or missing. Apart from mortal hazards, the decision of whether to try traveling by legal means is also important in determining the potential risks associated with a route.

Intermediate Destinations

During their journey, refugees may often stay, perhaps for prolonged periods, in another country along the way. In some circumstances, this will be among the general population, residing either with or without legal authorization. Alternatively, this may involve a stay in a designated refugee camp for periods as short as a few days or as long as a number of years. It is difficult to find reliable information about how typical it is for refugees to have had some experience in camps but clearly many arrive without ever having done so.

The UN High Commissioner for Refugees (2019a) estimated that 60 percent of refugees lived in noncamp accommodation in 2018, though of course this number varies widely from many developed countries, where essentially all refugees live in private accommodation, to some of the least developed countries where the majority of refugees reside in camps. Refugee camps vary greatly in their size, funding level, organization, and longevity, from Kutupalong in Bangladesh, established in 1991 and recently expanded to a population of over half a million, to La Linière in France, opened in 2016 and closed just a year later, housing 1,600 refugees at its peak. While it is difficult to generalize, refugee camp facilities are mostly rudimentary, opportunities for work and education are minimal or informal, and health and safety risks are common. Spending extended periods in a refugee camp could seriously

affect future prospects for integration into a developed labor market, because there may be limited opportunities to engage in the formal workforce while residing in a camp, and so residents' human capital may degrade over time.

A refugee camp may be a direct pathway to resettlement in a developed country, but this experience is not especially common (Hatton forthcoming): the UN High Commissioner for Refugees (2019a) records that only 92,400 refugees were resettled by 25 countries in 2018. Resettlement is one of three durable solutions considered by the UN High Commissioner for Refugees (2011) for refugees, voluntary repatriation or local integration being alternative possibilities. The process of selection for resettlement introduces a further set of criteria bearing on selection of the refugee population arriving in high-income countries. Of refugees that are not resettled, some will eventually decide to move on or return home, but many others may remain. Some long-standing camps have turned into de facto permanent towns or merged into nearby cities (such as Deir al-Balah in Gaza).

Arrival

The method of arrival in a host country, whether resettled, legally arriving directly, or illegally arriving, may have important implications for an asylum seeker's legal status and hence ability to undertake work. Resettled individuals will arrive with asylum status already determined and may therefore be at an advantage in joining the local labor market. Irregular arrivals, on the other hand, may be more likely to spend time in detention while their claims are being processed, which could have impacts on mental health as well as human capital. Of course, this is likely to vary significantly between host countries and over time as their policies change.

The nature of reception in the receiving country is also likely to be of great significance. Refugee status is not typically granted immediately and refugee migrants can find themselves subject to procedures of validation that inhibit their ability to work and aggravate feelings of alienation, perhaps even appearing to replicate experiences of interrogation and incarceration from which the individual may be fleeing (Phillimore 2011). Such procedures may hinder early labor market attachment, allowing skills to atrophy while the individual is unable to work, and create habitual persistence of dependence on welfare.

Furthermore, refugees are frequently subject to policies of forced dispersal, as described below for several north European countries, which isolate them from the sorts of social networks of previous immigrants that may be critical to job finding and social learning among typical migrants. In addition, refugees' integration and assimilation may be significantly hindered if they face hostility or discrimination from host communities.

To summarize, the labor market integration of refugees is likely more challenging than that of economically motivated migrants. We may expect refugees to arrive with skills less adapted to the receiving country's economic needs and to be of a composition that is less conducive—on average—to self-sufficiency through economic activity. Length and uncertainty of expected immigration duration may lead to conflicting effects on investment in skills specific to the receiving country's economy. Refugees are likely to be initially less well equipped with

productivity-enhancing proficiencies in host countries' labor markets and thus disadvantaged in comparison to economic migrants in terms of employment and wages. In the next section, we investigate whether this is borne out in the data.

Evidence on Labor Market Integration

Our investigation of the labor market integration of refugees focuses on employment and wages. One challenge in studying refugees is that they typically make up only a small fraction of the overall immigrant population, so that their numbers are small in general survey data. Moreover, most surveys or administrative datasets do not provide markers that allow a distinction to be drawn between economic and refugee migrants. Even when available, differences in measurement across receiving countries and differences in the definition of refugees mean that cross-country comparisons must be read with caution. In addition, refugees in different countries are subject to quite different integration policies and legal regimes, as well as often being drawn from quite different areas and cohorts. Disentangling these effects would be a challenge even with plentiful data.

Our analysis draws on three sets of data sources. First, we use various micro datasets that either focus specifically on refugees (including the UK's Survey of New Refugees and the Australian Building a New Life in Australia survey), contain refugee "boost" samples (the German Socio-Economic Panel), or that are detailed enough to naturally contain a meaningfully sized sample of the refugee population. Where data is from a publicly available survey covering only one country, we will refer to these as the "country-specific public survey" data. Second, also within the class of public survey data, we single out the EU Labour Force Survey (LFS), from which we use data collected during ad hoc modules administered in 2008 and 2014 that allow the identification of different types of immigrants, as a cross-national public survey. Finally, we have obtained from the authors of various papers on refugees that are based on census and register data, statistics on refugees and other immigrants' outcomes that will allow comparison across these countries.[4] We refer to these sources of data as the "administrative" data sources.

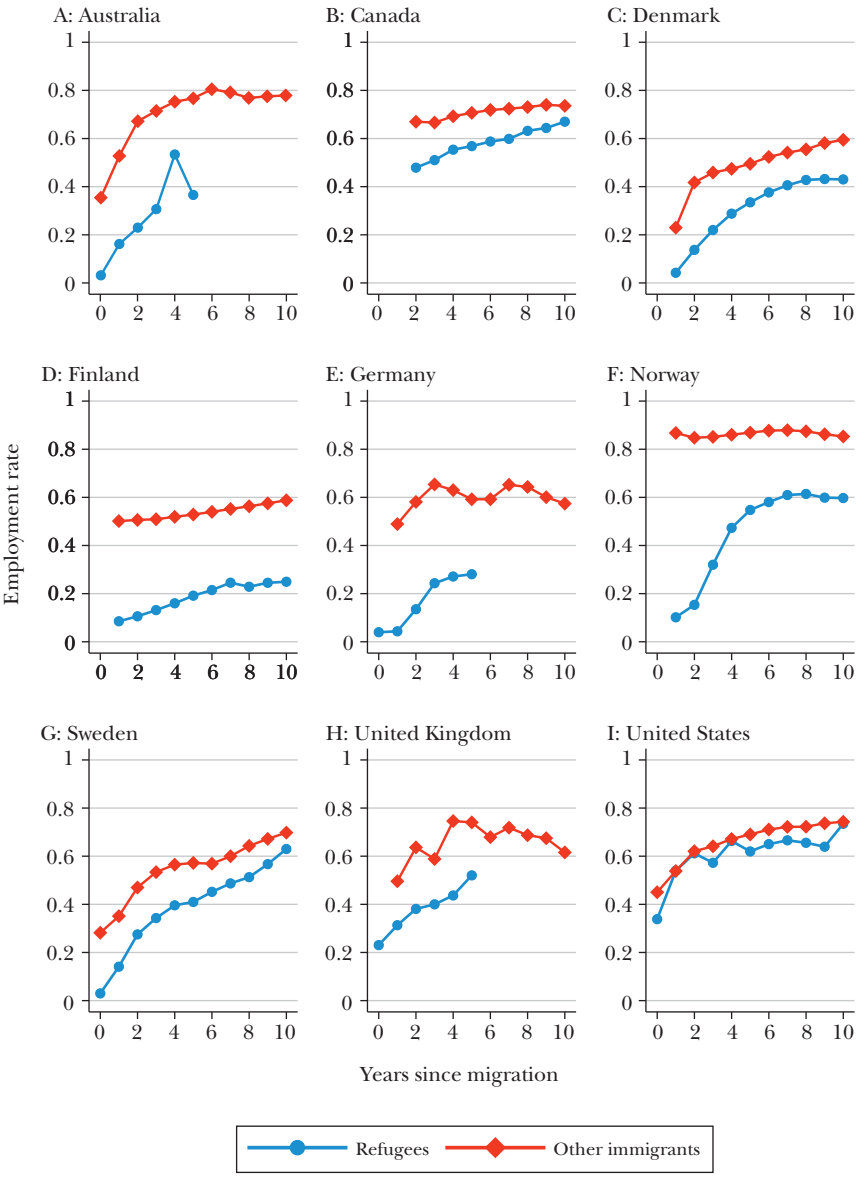
Each of these types of data has advantages and disadvantages, and we hope that—by providing evidence based on all three—we will be able to paint a comprehensive picture of the way in which refugees integrate into the labor markets of various countries, in comparison with other immigrants and natives.

Employment

Overall, employment rates of refugee migrants are very low immediately after arrival in the host country, but typically increase quite rapidly over the first few years after migration. However, there is significant heterogeneity between countries. Figure 2, drawing on administrative data and country-specific public survey

[4]These papers include Bevelander (2016); Bratsberg, Raaum, and Røed (2019); Mata and Pendakur (2017); Sarvimäki (2017); and Schultz-Nielsen (2017).

Figure 2
Employment Rates of Immigrant Groups over Time since Migration



Source: The results are based on data from the following sources (for details see the online Appendix): Australia—BNLA, HILDA; Canada—Census; Denmark—Administrative registers; Finland—Administrative registers; Germany—SOEP; Norway—Administrative registers; Sweden—Administrative registers; United Kingdom—SNR, LFS; and United States—ACS.
Note: The figure plots observed employment rates of refugees and other immigrants in various host countries over time after migration. The precise sample groups vary in their construction due to having been obtained from different data sources (see the online Appendix), but generally consist of working-age males and females.

datasets, shows the employment rates of refugees and other migrants (typically those who migrated for labor market and/or family related reasons) over time after migration for several host countries. Care should be taken when reading this plot, as the "other immigrant" samples vary in their construction and may not be precisely comparable to the refugee samples, but the general trends are clear.[5]

Except for the United Kingdom, the United States, and Canada, employment rates for refugees are below 20 percent in the first two years after arrival. In contrast, other immigrants have higher employment rates at arrival in all countries, though these still vary significantly between countries. The employment of refugees increases in subsequent years at different rates across countries: rapidly so in Australia, Sweden, and Norway, but more modestly in Denmark, Germany, and Finland. In some countries, such as Sweden and Canada, refugees appear to mostly close the employment gap with other immigrants after a decade in the country, while in others such as Norway and Finland, the gap remains large and stable over this period. The most notable outlier country in this figure is the United States, where refugees' employment rates track those of other immigrants closely. It is not entirely clear why the US experience appears so different in this figure; possible explanations could relate to the nature of the US labor market or to the nature of the settlement process in the United States, but require further investigation.

To complement Figure 2, the employment rates of refugees two years and ten years after migration are also listed in Table 1, along with the differences between the employment rates of refugees and natives and between refugees and immigrants with the same length of residency. For almost all countries, the gap between refugees and other groups is closing over time, although refugees have persistently lower employment rates than other immigrants and natives ten years after migration. As mentioned, the exception is the United States, where refugees appear to have caught up to other immigrants after just two years and to natives by ten years after migration (a finding that is compatible with the existing literature).

[5]In an online Appendix, we describe our sources and methodology in detail. Sources, samples, and empirical methods differ from series to series, and the "other immigrant" categories vary in their composition. Data sources include the Household, Income and Labour Dynamics in Australia (HILDA) survey (Department of Social Services and Melbourne Institute of Applied Economic and Social Research 2001–2017), the Building a New Life in Australia (BNLA) survey (Department of Social Services and Australian Institute of Family Studies 2013–2014), the German Socio-Economic Panel (SOEP) (German Institute for Economic Research 1984–2017; Goebel et al. 2019), the UK Labour Force Survey (LFS) (Office for National Statistics, Social Survey Division, Northern Ireland Statistics and Research Agency, and Central Survey Unit 2008), the UK's Survey of New Refugees (SNR) (Home Office, UK Border Agency: Analysis, Research and Knowledge Management 2010), the American Community Survey (ACS) (Ruggles et al. 2019), the US Yearbook of Immigration Statistics (YIS) (Office of Immigration Statistics 2001–2017), and the EU Labor Force Survey (LFS) (European Commission 2008; 2014). It should also be noted that some of the series presented are based on single cross sections, while others are drawn from longitudinal or repeated cross-sectional data. In those series based on single cross sections, variation over time since arrival is provided purely by analysis of different arrival cohorts, whereas for data covering multiple years of observation, changing outcomes over time of fixed cohorts are combined with variation between cohorts to give the overall effect. In both cases, selective outmigration plays a role in determining the observed composition of migrants who have been in the country a given number of years (Dustmann and Görlach 2015).

Table 1
Employment Outcomes of Refugees Compared to Other Groups

Host country	*Years since migration*	*Refugee employment rate*	*Gap to other immigrant employment rate*	*Gap to native employment rate*
Australia	2	0.23	0.44	0.55
Canada	2	0.48	0.19	0.27
Finland	2	0.11	0.40	0.64
Germany	2	0.14	0.45	0.57
Norway	2	0.15	0.69	0.73
Sweden	2	0.28	0.20	0.54
United Kingdom	2	0.38	0.26	0.38
United States	2	0.61	0.01	0.11
Canada	10	0.67	0.07	0.08
Finland	10	0.25	0.34	0.50
Norway	10	0.60	0.26	0.29
Sweden	10	0.63	0.07	0.19
United States	10	0.73	0.01	-0.01

Source: The results are based on data from the following sources (for details see the online Appendix): Australia—BNLA, HILDA; Canada—Census; Finland—Administrative registers; Germany—SOEP; Norway—Administrative registers; Sweden—Administrative registers; UK—SNR, LFS; and USA—ACS.
Note: The table compares observed refugee employment rates to those of other immigrants and natives for various host countries at two and ten years after migration to the country. The fourth and fifth columns show the amount by which the refugee employment rate trails that of other immigrants or natives, respectively. The precise sample groups vary in their construction due to having been obtained from different data sources (see the online Appendix), but generally consist of working-age males and females.

Table 2 provides additional detail, by distinguishing between employment growth rates over the first 5 years in the country and in years 6–10. On average, employment growth of refugees is substantially higher than that of other migrant groups in both periods, a regularity that also holds for almost all countries when viewed in isolation. Notably, while employment of other immigrants is close to flat for several countries in the second period, refugees continue to experience growth, indicating an integration process of longer duration.

A similar picture emerges from Figure 3, based on data instead from the 2014 EU Labour Force Survey. The figure plots the employment rate of refugees against that of other immigrants, for those who have been in the country for less than 10 years, between 10 and 19 years, and for more than 19 years. Each point represents a European country. The figure shows that for those who migrated less than a decade ago, refugees in almost every country experience substantially worse employment rates than other immigrants (the only exception being Switzerland), mirroring the findings from Figure 2 and Table 1. However, refugees with between 11 and 19 years residency are employed at rates much closer to other immigrants, and any difference appears to be largely erased for those with residency longer than 20 years.

Because the integration process may differ substantially for different demographic subgroups, we also considered employment outcomes of male and female groups separately. Refugee women appear to be employed at particularly low rates—the ratio of female to male employment rates is smaller for refugees than for other

Table 2

Employment Growth Rates of Refugees and Other Immigrants over Time since Arrival

Host country	*Refugees 0–5 years*	*Other immigrants 0–5 years*	*Refugees 5–10 years*	*Other immigrants 5–10 years*
Australia	0.067	0.083	—	—
Canada	0.030	0.012	0.020	0.006
Denmark	0.073	0.066	0.019	0.020
Finland	0.027	0.007	0.012	0.012
Germany	0.048	0.026	—	—
Norway	0.111	0.000	0.010	–0.003
Sweden	0.076	0.058	0.044	0.025
United Kingdom	0.058	0.061	—	—
United States	0.056	0.048	0.023	0.011
Average	0.061	0.040	0.021	0.012

Source: The results are based on data from the following sources (for details see the online Appendix): Australia—BNLA, HILDA; Canada—Census; Denmark—Administrative registers; Finland—Administrative registers; Germany—SOEP; Norway—Administrative registers; Sweden—Administrative registers; United Kingdom—SNR, LFS; and United States—ACS.

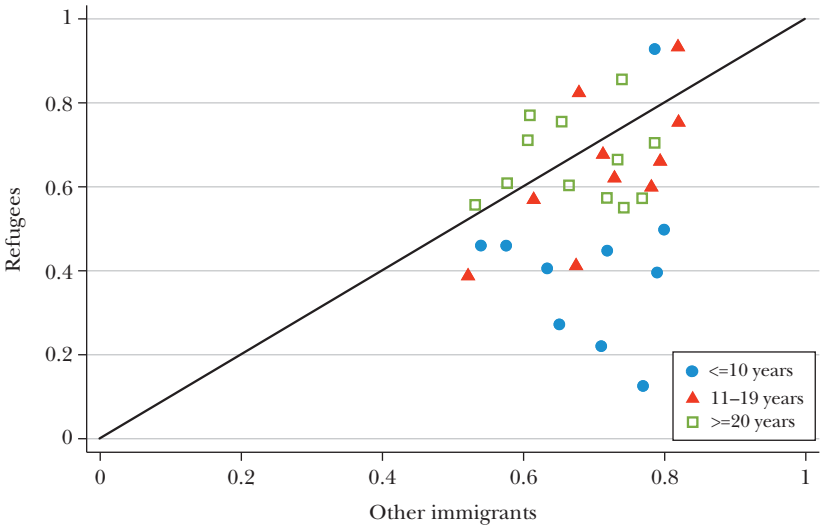
Note: The table shows average growth of employment rates for refugees and other immigrants. The second column shows the average yearly increase in the refugee employment rate observed during the first five years of residency in the host country, and the analogous figures for nonrefugee immigrants are displayed in the third column. The fourth and fifth columns similarly show the average yearly increases in employment observed for refugees and other immigrants during the period between five and ten years after arrival in the host country. The precise sample groups vary in their construction due to having been obtained from different data sources (see the online Appendix), but generally consist of working-age males and females.

immigrants (and both are smaller than for natives) in each country considered. This pattern is especially dramatic in the immediate years after migration, and while this ratio for refugees remains persistently smaller than that of natives even after a decade, in most cases, the difference between refugees and other immigrants appears to shrink significantly over this time scale. We also looked at the data across the countries in the EU Labour Force Survey to probe whether patterns of age, gender, or education level could explain some of the gaps we have seen between the outcomes of refugees and other immigrants. However, employment gaps conditional on these factors are qualitatively similar to the analogous unconditional results, leading us to the conclusion that differences in the demographic compositions of groups (at least in these dimensions) are not the main drivers of the differentials we have observed.[6]

Some general conclusions emerge from this discussion. First, initial employment rates of refugees are considerably lower than those of other immigrant groups. This finding is in line with our expectations, as refugees are likely to arrive with skills less adapted to the receiving country's labor market. Second, refugee employment

[6]For more detail on gender breakdowns and conditional labor market outcomes, see the online Appendix. The conditional employment plots are based on linear probability regressions, where we control for age, gender, and education.

Figure 3
Employment Rates of Immigrant Groups across European Countries



Source: This plot is based on data from the 2014 ad hoc module of the EU Labour Force Survey.
Note: This figure shows the employment rates of refugees compared to those of other immigrants for various European countries. Refugees are identified as those whose reported reasons for migration are international protection or asylum. The "other immigrants" sample consists of all other non-natives. Both groups are restricted to individuals between the ages of 20 and 64 whose main activity is not education or training (see the online Appendix for details). Each point in this figure represents a country, and the distance below the 45° line represents the extent to which refugees are employed at lower rates than other immigrants. This is shown separately for migrants who have been in the host country at most 10 years, between 11 and 19 years, and at least 20 years. Due to the small numbers of refugees in each individual country, some of the plotted points are calculated based on a small number of observations. Any individual point should be regarded as having limited reliability, though the general pattern can be expected to be more robust.

increases most sharply during the first two or three years after arrival. This pattern suggests that the first years after arrival are a crucial period for integration. Third, refugee employment continues to grow quickly for the rest of the first half-decade after the first few years and indeed continues to grow in the second half-decade, although at a slower rate. This pattern highlights that the time scale of integration appears to be much longer for refugees than for other immigrants. Fourth, employment levels of refugees in the longer term (a decade after arrival) continue to vary significantly between countries, but in many cases do not approach the levels of natives or other immigrants. However, there is some evidence that after the first decade, employment rates of refugees seem to converge to those of other immigrants. Finally, female refugees experience persistently lower employment rates than their male counterparts, and they are particularly missing out on the rapid employment growth experienced by men in the early years after migration (this is illustrated in online Appendix Figure A1).

Wages

In addition to being employed at lower rates than natives and other immigrants, even those refugees who do manage to find employment generally experience lower wages than the other groups. Their relative wage position gradually improves over time compared to an average native but not, in most countries, markedly faster than other immigrants. Again drawing on country-specific public survey and administrative data (reliable wage data being available only for a subset of countries for which we observe employment), we show in Figure 4 the average wage levels (calculated conditional on being in employment) of refugees and other immigrants as a fraction of average natives' wages over the first ten years after arrival.[7] In addition, we list average wage ratios of refugees and other groups after two and ten years in Table 3. For instance, while average wages of refugees who had been in the United States for two years amounted to 40 percent of native wages and 49 percent of other immigrants' average wages, after 10 years, average wages had improved to 55 percent of natives and 70 percent of other immigrants in the same position. It should be noted that changes in relative wages may be due to both wage changes of those in employment and changes in the composition of refugees who are in work.

Several general observations follow from Figure 4. First, as compared to employment rates where the growth in the first few years is much more rapid than that of subsequent years, refugee wages increase slowly but consistently relative to those of natives over time. Second, even in the long term, refugee wages often do not approach those of natives and continue to lag significantly behind those of other immigrants. Third, even in countries where refugee employment rates quickly approach the levels experienced by natives or other immigrants (like the United States), the corresponding wage gaps can remain large and persistent. Finally, while cross-country variation in refugee wages relative to natives is still significant, it is not nearly so large as that of employment rates.

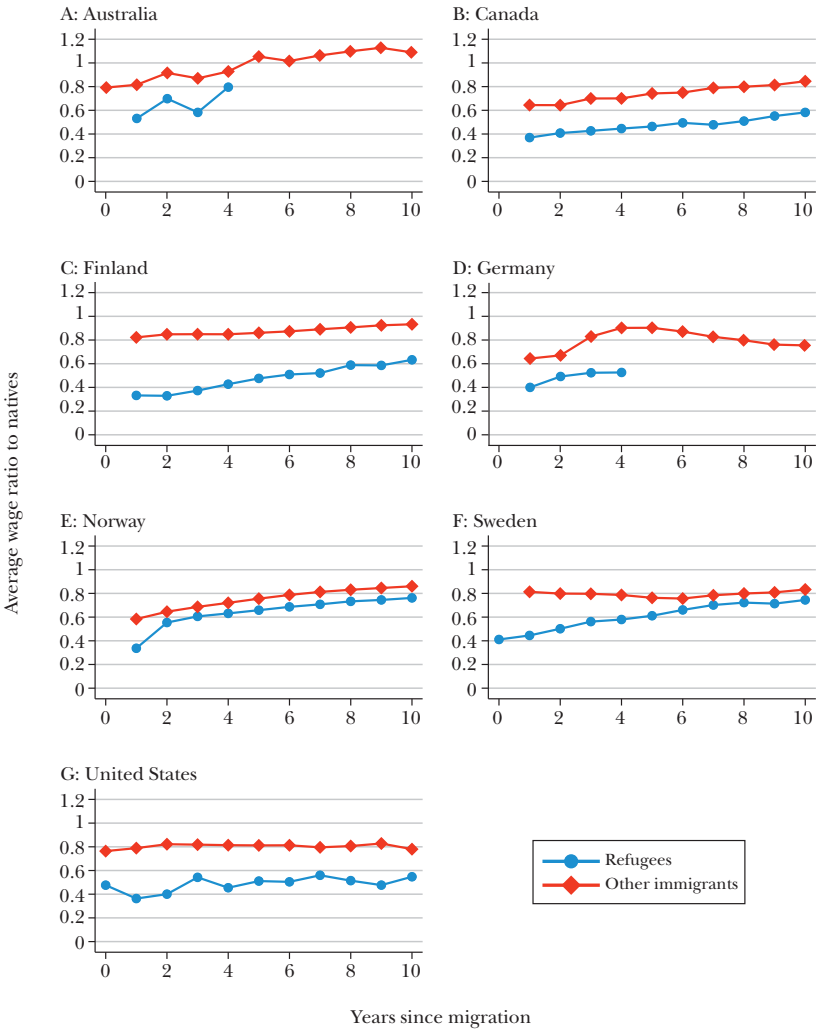
As with employment rates, we also investigated whether these results can be explained by compositional differences between refugee populations and other groups using data from the EU Labour Force Survey. Again, we do not see qualitative changes in the results when controlling for age, education, and gender, indicating that these factors are not the primary cause of the observed trends in refugee wages.

Previous Evidence

Overall, the patterns of refugee employment and wages discussed in the previous sections are consistent with the findings of previous literature. Of course,

[7] We simply calculate the average wage of all employed working-age natives without allowance for differences in age or other compositional factors and compare it to the average wage of all working-age refugees who have been in the country for a given number of years (and similarly for other migrants). The number of countries represented is fewer than in Figure 2, since we do not have reliable wage data for as many countries as we do for employment.

Figure 4
Wage Levels of Migrant Groups Compared to Natives over Time since Migration



Source: The results are based on data from the following sources (for details see the online Appendix): Australia—BNLA, HILDA; Canada—Census; Finland—Administrative registers; Germany—SOEP; Norway—Administrative registers; Sweden—Administrative registers; and United States—ACS.
Note: The figure plots the mean wages of immigrant groups (conditional on employment) in various host countries over time after migration. These wages are presented as a fraction of the mean wages of the native population. The precise sample groups vary in their construction due to having been obtained from different data sources (see the online Appendix), but generally consist of working-age persons.

other studies also offer different areas of focus and thus can fill in some other facets of the picture. For a review of the literature on immigrant integration, De la Rica, Glitz, and Ortega (2015) offers a useful starting point. Dustmann and Görlach (2015) provide an assessment of the empirical challenges in estimating earnings

Table 3
Wages of Refugees Compared to Other Groups

Host country	*Years since migration*	*Refugee to native wage ratio*	*Refugee to other immigrant wage ratio*
Australia	2	0.697	0.761
Canada	2	0.408	0.634
Finland	2	0.329	0.388
Germany	2	0.496	0.735
Norway	2	0.554	0.858
Sweden	2	0.502	0.628
United States	2	0.401	0.487
Canada	10	0.583	0.689
Finland	10	0.633	0.678
Norway	10	0.762	0.886
Sweden	10	0.745	0.894
United States	10	0.547	0.701

Source: The results are based on data from the following sources (for details see the online Appendix): Australia—BNLA, HILDA; Canada—Census; Finland—Administrative registers; Germany—SOEP; Norway—Administrative registers; Sweden—Administrative registers; and United States—ACS.
Note: The table compares average wage levels of employed refugees to those of other immigrants and natives for various host countries at two and ten years after migration to the country. The third and fourth columns show the ratio of refugee wages to natives and other immigrants, respectively. The precise sample groups vary in their construction due to having been obtained from different data sources (see the online Appendix), but generally consist of working-age males and females recorded as being in employment.

assimilation for immigrant populations. Less is known about the economic integration of refugee immigrants specifically, though a substantial literature has begun to develop in recent years. For reviews of the existing evidence on refugee labor market integration, useful starting points are Chin and Cortes (2015), Bevelander (2016), and Becker and Ferrara (2019).

For the United States, the previous literature suggests that refugees' employment rates are not dissimilar to those of other immigrants, but a large initial gap in earnings exists, with a subsequent relative improvement. For example, Cortes (2004) broke ground by looking at refugees together with, but distinguished from, other immigrants. Using public-use census data from 1980 and 1990, she separated immigrants arriving between 1975 and 1980 into refugees and economic immigrants according to country of origin and year of immigration. Refugees are found to initially earn less and work fewer hours than other immigrants, but their earnings grow faster. The difference between the groups is attributed to longer expected duration of stay. Chin and Cortes (2015) show how this steeper path of labor market outcomes is associated with greater gains in education and language proficiency.

Studies have also looked at occupational prestige or status, which attempts to measure the extent to which, say, a refugee who is an engineer or teacher in another country may end up driving a cab or working in a fast-food restaurant in a high-income country. Akresh (2008) used survey data from the 2003 New Immigrant Survey, which records the last job held abroad, to show that refugees display

the sharpest downgrading in occupational prestige and the steepest subsequent upgrading of any immigrant group. Using the same survey, Connor (2010) shows that refugees, while employed at similar rates to other immigrants, still suffer a gap in earnings and occupational status, attributable in large part to differences in education, language ability, and neighborhood.

Both the time at which refugees arrive and their age at arrival can affect their integration prospects as well. Capps et al. (2015) and Fix, Hooper, and Zong (2017) document more recent outcomes using the American Community Survey, identifying refugees indirectly by country of origin and year of arrival and showing refugees continuing to lag behind natives in incomes and education, but not employment rates. Evans and Fitzgerald (2017) use the same approach and data and focus on the importance of age at arrival. Refugees arriving in the United States before age 14 perform similarly to natives, teenage entrants do somewhat worse, and adult refugees do much worse in employment, earnings, and welfare dependency (though there is rapid improvement in early years).

In contrast to the US experience, refugees in European countries seem to lag behind other immigrants not just in earnings, but also in employment rates, although there is evidence for some catch-up in both dimensions over time. The European evidence seems to also be mirrored by studies for Canada (Aydemir 2011; Bevelander and Pendakur 2014), which tell stories of initial disadvantage but rapid growth in employment rates for refugees.

For Europe, a concentration of papers based on excellent register data investigate the labor market integration of refugees for Scandinavian countries.[8] Unlike the situation in the United States, refugees in these countries are observed to experience very low employment rates in the initial years after migration. Although their position improves during the first decade in the country, they typically do not close the gap to natives and other immigrant groups and even sometimes appear to fall away over time (Bratsberg, Raaum, and Røed 2014, 2017; Schultz-Nielsen 2017). Low labor market attachment leads to high welfare dependence observed in these studies. Among those who are employed, earnings are low (Schultz-Nielsen 2017; Sarvimäki 2017; Bratsberg, Raaum, and Røed 2014, 2017), though earnings trajectories are steeper for refugees than for other migrant groups (Bevelander 2011, 2016). Local employment conditions matter, particularly for the low-skilled (Bevelander and Lundh 2007), and integration patterns are different for different origin groups (Lundborg 2013). Bakker, Dagevos, and Engbersen (2017) provide an example from the Netherlands of the use of register data elsewhere in Europe, finding again that refugees begin at a large disadvantage compared to other immigrant groups, but that the gap closes over time.

Other analyses for European countries are typically based on survey data. The finding of large gaps in employment, income, and job quality relative to other migrants, which diminish over time, is confirmed by a number of papers using the EU

[8]For Denmark, see Schultz-Nielsen (2017); for Finland, see Sarvimäki (2017); for Norway, see Bratsberg et al. (2014, 2017); for Sweden, see Åslund, Forslund, and Liljeberg (2017), Bevelander and Lundh (2007), Bevelander and Pendakur (2009, 2014), Bevelander (2011), and Lundborg (2013).

Labour Force Survey, a large dataset with ad hoc modules on migrants in 2008 and 2014 (Dumont et al. 2016; Dustmann et al. 2017; Fasani, Frattini, and Minale 2018; Zwysen 2019).

For the United Kingdom, Bloch (2008) identifies high levels of overqualification among employed refugees. A number of papers (see the discussion in Ruiz and Vargas-Silva 2017, 2018) use the UK Labour Force Survey to show that refugees initially have lower employment and wages than comparable economic migrants but show faster growth, at least in employment. Ruiz and Vargas-Silva (2017) and Cebulla, Daniel, and Zurawan (2010) find similar results using the UK Survey of New Refugees.

Other Factors Affecting Refugee Labor Market Outcomes

There are many reasons why the labor market integration of refugees might be expected to differ from that of other migrants. The backgrounds and histories of refugees may inhibit labor market attachment or suppress the wages they can command in a host country. One potential mechanism is that both the selection of refugees and their experience of flight may mean that health status, and especially mental health status, will differ from both natives and other migrants.

For similar reasons, refugees' difficulties in economic integration are also expected to coincide with slower integration in broader social dimensions. After arrival, the development of host-country language skills and social networks are simple markers for social integration and will also clearly be important determinants of success in economic integration. We discuss these factors in this section, noting how refugees differ from other migrants and the resulting effect this is expected to have on labor market outcomes.

Health

Although many studies have found immigrants in general to be typically healthier at arrival than natives, refugees tend to arrive with lower levels of health than other types of immigrants (for example, Giuntella et al. 2018). For the United States, Chin and Cortes (2015) find refugees are almost twice as likely to report being in "poor" or "fair" health as compared to other immigrants (17 versus 9 percent) and similarly much more likely to report being "troubled by pain" (18 versus 9 percent). This difference could be both due to the fact that refugees are selected in a different way than other migrants (in particular, with lower human capital, which has a positive association with health) and due to the deleterious effects of their experiences in their home country or during their subsequent flight.

Fleeing traumatic and emotionally damaging circumstances will affect psychological and physical health, and occurrence of mental health difficulties among refugee populations is well evidenced (Porter and Haslam 2005). This may only aggravate the particularly low initial economic fitness and adaptability of refugees as recovery from trauma and continuing distress over the circumstances from which the individual has fled distracts from integration (for example, Phillimore 2011). In particular, the

incidence of mental illness among refugees is likely to be much higher than in the general population, due to experiences of violent, life-threatening, and traumatizing events in their origin country, adverse conditions during flight or in refugee camps, and potentially exposure to violence or sexual and physical exploitation during and after migration. In addition, stress and anxiety caused by uncertainty about their status in a host country can be expected to exacerbate these problems. Schock et al. (2016), studying refugees in Germany, report that more than 60 percent of adult refugees and more than 40 percent of adolescents have experienced violence in their countries of origin and/or during their migration. Mental health conditions may be an important factor that inhibits the ability of individuals to cope with an unfamiliar environment by disrupting the acquisition of new skills and establishment of social contacts. Indeed, some studies have found mental health indicators to be important predictors of refugee labor market outcomes: for example, in the Netherlands (De Vroome and van Tubergen 2010) and the United Kingdom (Ruiz and Vargas-Silva 2018).

Estimates on the prevalence of mental health disorders among refugees vary considerably, but the overall picture is quite clear of an alarming incidence of mental health issues, in particular depression and post-traumatic stress disorder (for example, Bogic, Njoku, and Priebe 2015; Priebe, Giacco, and El-Nagib 2016; Giacco, Laxhman, and Priebe 2018). Bogic, Njoku, and Priebe (2015) point out that around two-thirds of studies of longer term refugees (displaced for more than five years) report prevalence of post-traumatic stress disorder greater than 20 percent (although lower quality studies tended to report higher rates). Focusing on more reliable studies, the authors suggest that refugees may be several times more likely than general Western populations to suffer either from post-traumatic stress disorder or from depression.

Another possible consequence of refugees' traumatic or violent experiences, along with inhibiting their integration into the host society and economy, may be antisocial behavior after resettlement. Studying the relation between exposure to conflict and violent behavior of refugees in Switzerland, Couttenier et al. (2019) report that cohorts exposed to civil conflicts or mass killings during childhood are on average 40 percent more prone to violent crimes than conationals without this exposure. Moreover, the heterogeneity of integration policies across cantons also allows the authors to show that these effects can be eliminated through policies encouraging early labor market attachment. Horyniak et al. (2016) link trauma and mental illness among refugees, particularly men, to substance abuse.

Thus, the existing evidence seems to suggest that refugees' experiences with violence and trauma can have serious effects on their mental health, and that the share of refugees suffering mental illnesses such as post-traumatic stress disorder is far higher than that in the general populations of host countries. This in turn will have serious consequences for their labor market integration, as well as for the host society in general.

Language

Proficiency in the language of the receiving country is among the most salient and frequently discussed aspects of human capital deficiency among arriving

immigrants (for example, Dustmann and Fabbri 2003). In the United States, numerous authors have provided evidence of the initial weakness and formidable subsequent role of English fluency in adaptation of refugees to the US labor market (as in Connor 2010; Chin and Cortes 2015; Evans and Fitzgerald 2017).

In Europe, Dumont et al. (2016) document large variation between EU host countries in the levels of refugee language proficiency: for example, higher in Spain and lower in Germany. Across the European Union as a whole, 24 percent of refugees with less than ten-years residence have advanced host-country language knowledge, increasing to 49 percent for those with more than ten-years residence (whereas the analogous figures for other non-EU born are 54 percent and 69 percent, respectively). Indeed, much of the gap between native and refugee employment in the European Union is argued to be accounted for by differing language skills: 59 percent of refugees with at least intermediate-level host-country language skills are employed as opposed to only 27 percent of those below this level.

More directly addressing the mechanisms linking language proficiency and employment, Fasani, Frattini, and Minale (2018) report that about one-quarter of refugees across Europe cite language difficulties as the principal obstacle to employability and Bloch (2008) gives a similar figure for the United Kingdom. Auer (2018) uses random assignment of refugees across Swiss language regions as a plausible source of exogenous variation and finds an association of language knowledge with increased probability of job finding.

To demonstrate directly how language skills of refugees compare to those of other migrants and how this changes over time, we use the EU Labour Force Survey's 2014 ad hoc module on the labor market situation of migrants. Immigrants were asked to rate their proficiency in the host country's language from "beginner or less," "intermediate," "advanced," or "mother tongue."[9] The overall pattern is that refugees consistently appear to begin with lower language proficiency than other immigrants (the only exception being in Switzerland). While the language skills of both refugees and other migrant groups appear to improve slowly but substantially over time, refugees' proficiency seems to persistently lag behind that of the other immigrant groups, even decades after migration.

As with labor market outcomes, the story does, however, appear slightly different in the United States. Looking at the American Community Survey (ACS), language proficiency is recorded on a five-response scale from "does not speak English" to "speaks only English at home." The results of this survey again show that refugees arrive with lower levels of language proficiency than other migrants—at the time of migration, only about 44 percent of refugees speak English "well" or better, compared with 64 percent of other immigrants. However, while other immigrants do not tend to see particularly strong gains in English speaking skills over time, refugees rapidly improve and even overtake other migrants' speaking abilities around ten years after arriving in the United States.

[9]For more details on the evidence about language proficiency of refugees discussed throughout this section, the online Appendix offers more detail on language skills for refugees and other immigrants, including figures illustrating both the EU and the US data.

The American Community Survey also asks about linguistic isolation, measured by whether an individual lives in a household in which no person above the age of 14 speaks English "very well" or better. Refugees are initially much more likely than other immigrants to live in houses in which no member is proficient in English, by a margin of 54 percent to 32 percent. Again, while other immigrants do not see much change in this measure over two decades, refugees' rate of linguistic isolation rapidly drops in the years following migration, falling below that of other immigrants after around a decade. Together, these patterns suggest that considerable effort is made in the refugee population to acquire English language proficiency, seemingly above that of other US immigrant groups.

In addition to having well-documented impacts on employability and other economic outcomes, language proficiency is also more generally important for social integration. In particular, Cheung and Phillimore (2014) demonstrate its importance to social network formation.

Social Networks

The formation of social connections, including both bonds with conationals or co-ethnics and bridges to native communities, is important to the broader refugee integration process (Ager and Strang 2008; Cheung and Phillimore 2014) and assists in the economic assimilation of refugees. The economic literature typically measures social networks in an indirect way, by counting individuals of same or similar origin in the region of settlement. An obvious problem of inferring the economic effects of social networks arises if there is sorting—say, if newcomers are more likely to choose to settle where economic conditions are favorable. This concern is typically addressed in the literature by concentrating on situations of random settlement policies for refugees.

The existence of local social networks, as well as evidently being an important measure of social integration per se, has also been argued to be important for migrants' job search prospects—for example, if job opportunities are communicated through established networks such as ethnic communities. Beaman (2012) develops a model along these lines in which employed individuals pass job offers to unemployed network members. In the short run, new arrivals increase the number of unemployed individuals seeking job information, while the number of employed members who can provide this information remains unchanged, which implies that a surge of recently arrived refugees has a negative effect on job finding rates in the short term. However, as refugees do become employed and thus able to pass along additional job offers, a positive information effect eventually dominates. Examining these implications for the labor market outcomes of refugees resettled in the United States, Beaman finds that an increase in the number of social network members resettled in the same year or one year prior to a new arrival leads to a deterioration of outcomes, while a greater number of tenured network members improves the probability of employment and raises wages.

Evidence from Europe generally supports a similar story, with larger social networks improving the labor market outcomes of refugees. For example, making use of dispersal policies for refugees in Scandinavia, several authors (for

Sweden, Edin, Fredriksson, and Åslund 2003, 2004; for Denmark, Damm 2009, 2014) have found that living in areas with high concentrations of co-ethnic or other minority individuals can improve the labor market outcomes of these refugees. These studies find that the effects of larger social networks are amplified for members of higher skilled or better employed groups, which is consistent with Beaman's (2012) model of job information dissemination through ethnic networks. In line with these results, Brücker et al. (2019) find evidence that dispersal policies in Germany have harmful effects on the labor market outcomes of the dispersed refugees. Further supporting the story of job opportunity transmission through social networks, Dagnelie, Mayda, and Maystadt (2019) find evidence for refugees in the United States that employment probability is affected positively by the number of business owners and negatively by the number of employees in their network.

Overall, access to a larger social network of established previous migrants seems helpful in transmitting information and providing access to preferential employment possibilities for newly arrived refugees.

Discussion and Policy Implications

A substantial body of evidence paints a highly consistent picture of refugees as disadvantaged socially and economically relative to other immigrants at arrival. We have provided a comprehensive review of refugees' economic integration and associated processes such as their social integration, language acquisition, and health outcomes, drawing together the existing literature and analyzing an inclusive collection of data from numerous sources and countries. Our focus has been on Europe, Australasia, and North America, regions that, despite a recent rise, receive only a fraction of the worldwide refugee population. Additional future analysis investigating similar issues for receiving countries outside this high-income group would be very timely.

Based on our investigation, we can conclude that refugees have—with the United States being an exception—substantially lower employment rates than other immigrants for at least the first decade after arrival, but that the gap comes close to disappearing during the second decade. Those refugees who do find work also experience much lower wages than other immigrants; again, the gap becomes smaller, but does not close during the first decade. The gap in labor market achievement between refugees and other immigrant groups (and indeed natives) is mostly unaccounted for by differences in demographic composition and the educational disadvantage of refugee groups. Aggravating factors for the detrimental economic position of refugees could include language deficiencies or physical and mental health problems due to experiences in regions of origin or during migrations.

One area of reform that can facilitate early integration is the asylum process itself, which is often lengthy and unpleasant. An important finding from the existing literature is that the length of time spent in refugee camps or other asylum accommodation has a strong impact on the future outcomes of refugees. For instance, for the Netherlands, Bakker, Dagevos, and Engbersen (2014) find

that a longer stay in asylum accommodation decreases the likelihood and quality of future employment, while De Vroome and Van Tubergen (2010) establish a negative association between the time spent in refugee reception centers and economic integration. Hainmueller, Hangartner, and Lawrence (2016) show that for refugees in Switzerland, each additional year that an asylum seeker waits for their claim to be processed decreases the subsequent employment rate by several percentage points. Similarly, Hvidtfeldt et al. (2018) compute that an additional year of waiting time in the Danish asylum system decreases subsequent employment by 3.2 percentage points on average. Hvidtfeldt et al. (forthcoming) show that lengthened waiting times also raise the risk of psychiatric problems. In Germany, Brücker et al. (2019) find that prolonged asylum procedures inhibit subsequent job finding.

Asylum claims may be decided while outside the country of ultimate destination, possibly in camps near to the origin country, or may be decided after arrival in the potential host country, but while still living in restricted housing conditions with barriers to employment and while supported by state payments. These barriers may have effects that persist long after the formal restrictions are lifted. Marbach, Hainmueller, and Hangartner (2018) show that temporary employment bans after arrival in Germany have significant adverse effects on subsequent employment trajectories of refugees.

After acceptance of refugees, it is not uncommon for host countries to enforce regional dispersal. The general argument for these policies is that this spreads the burden of support, avoids enclaves, forces refugees to engage with receiving communities, and therefore incentivizes acquisition of human capital and accelerates integration. However, the evidence suggests that if economic integration is the objective, this approach is questionable. Dispersal of refugees means depriving them of access to networks of individuals of similar origin, which are often critical to job finding and social learning. Thus, allowing for unrestricted settlement decisions of refugees within the receiving country may lead to better economic outcomes than external allocation.

In terms of post-arrival policy choices that can improve refugees' mental health outcomes, the German National Academy of Sciences Leopoldina (2018), in a detailed analysis of the various channels through which experiences of refugees can affect their mental health, emphasize the importance of providing support addressing psychological problems at an early stage. Giacco, Laxhmant, and Priebe (2018), as well as several other studies, emphasize the detrimental and aggravating effects that adverse conditions in a host country can have on refugees' mental health. Similar conclusions are reached by Bakker, Dagevos, and Engbersen (2014) and Kaltenbach et al. (2018), while Porter and Haslam (2005) identify living in institutional accommodation and experiencing restricted economic opportunity as risk factors for mental health outcomes. Studies investigating mental health outcomes in relation to post-migration experiences overwhelmingly conclude that the consequences of exposure to violence and trauma can be mitigated by early psychological support, reduced duration in asylum facilities, and support for early absorption into the labor market.

We conclude therefore that keeping the asylum process short, providing early support to address health issues, and facilitating refugees to join the labor market at the earliest possible stage are of key importance. Such policies reduce skill loss, help to reduce uncertainty about future residence, and improve the effectiveness of human capital investment, thus enhancing incentives to invest. To underscore this point, Bakker, Dagevos, and Engbersen (2014) find that in the Netherlands, temporary legal status leads to lower employment probability and job quality than permanent legal status and naturalization. Fasani, Frattini, and Minale (2018) show that groups of refugees granted permanent status at higher rates experience more favorable labor market outcomes. The success of such policies is also consistent with the earlier evidence on economic integration, which suggests large initial skill deficiencies that can potentially be addressed by policy.

Over and above all of this, refugees may find themselves subject to particularly intense hostility from host communities suspicious of the genuineness of claims of persecution and influenced by populist campaigns portraying asylum seekers as opportunistic exploiters of misplaced generosity. Public policy can accentuate or ameliorate such hostilities, at least to some extent.

In coming years, the outflow of refugees from poorer regions of the world seems likely to continue undiminished, given the continued political fragility of populous and growing countries from which migration to safer locations is increasingly easy. International obligations mandate a humanitarian duty to provide refuge in well-established cases. Reluctant acceptance of those obligations with arduous asylum processes and conditions that hinder successful integration harms the interests of refugees, wasting their talents and therefore also harming receiving countries themselves. A deeper understanding of the refugee experience can help to support sensible and constructive integration policy that encourages economically and socially productive participation of refugees in receiving societies.

■ *We are very grateful to Pieter Bevelander, Bernt Bratsberg, Ravi Pendakur, Matti Sarvimäki, and Marie Louise Schultz-Nielsen for making moments from their data on refugee integration available to us. This paper is in part based on data from Eurostat, Labour Force Survey, 2008 and 2014. This paper uses unit record data from the Household, Income and Labour Dynamics in Australia Survey (HILDA) and Building a New Life in Australia: the Longitudinal Study of Humanitarian Migrants (BNLA) conducted by the Australian Government Department of Social Services (DSS). The findings and views reported in this paper, however, are those of the authors and should not be attributed to the Australian Government, DSS, or any of DSS' contractors or partners. The responsibility for all conclusions drawn from the data lies entirely with the authors. Christian Dustmann acknowledges financial support from the DFG (grant number DU 1024/1-2 AOBJ: 642097) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement Number 833861). We are grateful also for comments from and discussion with Herbert Brücker, Anna Piil Damm, Francesco Fasani, Tommaso Frattini, Jens Hainmueller, Gordon Hanson, Tim Hatton, Marie Louise Schultz-Nielsen, Timothy Taylor, and Heidi Williams.*

References

Adams, Ben, and Cassandra Vinograd. 2015. "Salma's Story: One Refugee Family's Journey from Syria to Germany." *NBC News*, October 7. https://www.nbcnews.com/storyline/europes-border-crisis/refugee-crisis-one-familys-journey-syria-germany-n425636.

Adda, Jerome, Christian Dustmann, and Joseph-Simon Gorlach. 2019. "The Dynamics of Return Migration, Human Capital Accumulation, and Wage Assimilation." Unpublished.

Ager, Alastair, and Alison Strang. 2008. "Understanding Integration: A Conceptual Framework." *Journal of Refugee Studies* 21 (2): 166–91.

Akresh, Ilana Redstone. 2008. "Occupational Trajectories of Legal US Immigrants: Downgrading and Recovery." *Population and Development Review* 34 (3): 435–56.

Aksoy, Cevat Giray, and Panu Poutvaara. 2019. "Refugees' and Irregular Migrants' Self-Selection into Europe: Who Migrates Where?" CESifo Working Paper Series 7781.

Alcorn, Gay. 2019. "'The Land Where We Lived Has Gone'—The Life Story of a Rohingya Refugee." *The Guardian*, August 4. https://www.theguardian.com/world/2019/aug/04/rohingya-refugee-myanmar-australia-oppression-suffering.

Åslund, Olof, Anders Forslund, and Linus Liljeberg. 2017. "Labour Market Entry of Non-labour Migrants—Swedish Evidence." *Nordic Economic Policy Review* 2017: 115–58.

Auer, Daniel. 2018. "Language Roulette—The Effect of Random Placement on Refugees' Labour Market Integration." *Journal of Ethnic and Migration Studies* 44 (3): 341–62.

Aydemir, Abdurrahman. 2011. "Immigrant Selection and Short-Term Labor Market Outcomes by Visa Category." *Journal of Population Economics* 24 (2): 451–75.

Bakker, Linda, Jaco Dagevos, and Godfried Engbersen. 2014. "The Importance of Resources and Security in the Socio-economic Integration of Refugees. A Study on the Impact of Length of Stay in Asylum Accommodation and Residence Status on Socio-economic Integration for the Four Largest Refugee Groups in the Netherlands." *Journal of International Migration and Integration* 15 (3): 431–48.

Bakker, Linda, Jaco Dagevos, and Godfried Engbersen. 2017. "Explaining the Refugee Gap: A Longitudinal Study on Labour Market Participation of Refugees in the Netherlands." *Journal of Ethnic and Migration Studies* 43 (11): 1775–91.

Beaman, Lori A. 2012. "Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the US." *Review of Economic Studies* 79 (1): 128–61.

Becker, Sarah O., and Andreas Ferrara. 2019. "Consequences of Forced Migration: A Survey of Recent Findings." *Labour Economics* 59: 1–16.

Bevelander, Pieter. 2011. "The Employment Integration of Resettled Refugees, Asylum Claimants, and Family Reunion Migrants in Sweden." *Refugee Survey Quarterly* 30 (1): 22–43.

Bevelander, Pieter. 2016. "Integrating Refugees into Labor Markets." IZA World of Labor 2016: 269.

Bevelander, Pieter, and Christer Lundh. 2007. "Employment Integration of Refugees: The Influence of Local Factors on Refugee Job Opportunities in Sweden." IZA Discussion Paper 2551.

Bevelander, Pieter, and Ravi Pendakur. 2009. "The Employment Attachment of Resettled, Refugees, Refugees and Family Reunion Migrants in Sweden." In *Resettled and Included? The Employment Integration of Resettled Refugees in Sweden*, edited by Pieter Bevelander, Mirjam Hagström, and Sofia Rönnqvist, 227–45. Malmö, Sweden: Malmö University Electronic Publishing.

Bevelander, Pieter, and Ravi Pendakur. 2014. "The Labour Market Integration of Refugee and Family Reunion Immigrants: A Comparison of Outcomes in Canada and Sweden." *Journal of Ethnic and Migration Studies* 40 (5): 689–709.

Bloch, Alice. 2008. "Refugees in the UK Labour Market: The Conflict between Economic Integration and Policy-Led Labour Market Restriction." *Journal of Social Policy* 37 (1): 21–36.

Bogic, Marija, Anthony Njoku, and Stefan Priebe. 2015. "Long-Term Mental Health of War-Refugees: A Systematic Literature Review." *BMC International Health and Human Rights* 15: Article 29.

Bratsberg, Bernt, Oddbjørn Raaum, and Knut Røed. 2014. "Immigrants, Labour Market Performance and Social Insurance." *Economic Journal* 124 (580): F644–F683.

Bratsberg, Bernt, Oddbjørn Raaum, and Knut Røed. 2017. "Immigrant Labor Market Integration across Admission Classes." *Nordic Economic Policy Review* 2017: 17–54.

Bratsberg, Bernt, Oddbjørn Raaum, and Knut Røed. 2019. "Social Insurance Design and the Economic Integration of Immigrants." In *Integrating Immigrants into the Nordic Labour Markets*, edited by Lars Calmfors and Nora Sánchez Gassen, 133–58. Copenhagen, Denmark: Nordic Council of Ministers.

Brücker, Herbert, Jens Hainmueller, Dominik Hangartner, Philipp Jaschke, and Yuliya Kosyakova. 2019. "Refugee Migration to Germany Revisited: Some Lessons on the Integration of Asylum Seekers."

Paper presented at the XXI European Conference of the fRDB on "How to Manage the Refugee Crisis." Reggio Calabria, Italy, June 15.

Capps, Randy, Kathleen Newland, Susan Fratzke, Susanna Groves, Gregory Auclair, Michael Fix, and Margie McHugh. 2015. "The Integration Outcomes of U.S. Refugees: Successes and Challenges." Washington, DC: Migration Policy Institute.

Cebulla, Andreas, Megan Daniel, and Andrew Zurawan. 2010. "Spotlight on Refugee Integration: Findings from the Survey of New Refugees in the United Kingdom." London, UK: UK Home Office.

Cheung, Sin Yi, and Jenny Phillimore. 2014. "Refugees, Social Capital, and Labour Market Integration in the UK." *Sociology* 48 (3): 518–36.

Chin, Aimee, and Kalena E. Cortes. 2015. "The Refugee/Asylum Seeker." In *Handbook of the Economics of International Migration*, Vol. 1, edited by Barry R. Chiswick and Paul W. Miller, 585–658. Amsterdam, Netherlands: Elsevier.

Connor, Phillip. 2010. "Explaining the Refugee Gap: Economic Outcomes of Refugees Versus Other Immigrants." *Journal of Refugee Studies* 23 (3): 377–97.

Cortes, Kalena E. 2004. "Are Refugees Different from Economic Immigrants? Some Empirical Evidence on the Heterogeneity of Immigrant Groups in the United States." *Review of Economics and Statistics* 86 (2): 465–80.

Couttenier, Mathieu, Veronica Petrencu, Dominic Rohner, and Mathias Thoenig. 2019. "The Violent Legacy of Conflict: Evidence on Asylum Seekers, Crime, and Public Policy in Switzerland." *American Economic Review* 109 (12): 4378–425.

Dagnelie, Olivier, Anna Maria Mayda, and Jean-François Maystadt. 2019. "The Labor Market Integration of Refugees in the United States: Do Entrepreneurs in the Network Help?" *European Economic Review* 111: 257–72.

Damm, Anna Piil. 2009. "Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-experimental Evidence." *Journal of Labor Economics* 27 (2): 281–314.

Damm, Anna Piil. 2014. "Neighborhood Quality and Labor Market Outcomes: Evidence from Quasi-random Neighborhood Assignment of Immigrants." *Journal of Urban Economics* 79: 139–66.

De la Rica, Sara, Albrecht Glitz, and Francesc Ortega. 2015. "Immigration in Europe: Trends, Policies, and Empirical Evidence." In *Handbook of the Economics of International Migration*, Vol. 1, edited by Barry R. Chiswick and Paul W. Miller, 1303–62. Amsterdam, Netherlands: Elsevier.

Department of Social Services and Australian Institute of Family Studies. 2013–2014. "Building a New Life in Australia: The Longitudinal Study of Humanitarian Migrants, Release 4 (Waves 1–4)." Australian Data Archive Dataverse. doi:10.26193/ZQHBPW (accessed April 1, 2019).

Department of Social Services and Melbourne Institute of Applied Economic and Social Research. 2001–2017. "The Household, Income and Labour Dynamics in Australia (HILDA) Survey, GENERAL RELEASE 17 (Waves 1–17)." Australian Data Archive Dataverse. doi:10.26193/PTKLYP (accessed April 1, 2019).

De Vroome, Thomas, and Frank van Tubergen. 2010. "The Employment Experience of Refugees in the Netherlands." *International Migration Review* 44 (2): 376–403.

Dumont, Jean-Christophe, Thomas Liebig, Jorg Peschner, Filip Tanay, and Theodora Xenogiani. 2016. "How are Refugees Faring on the Labour Market in Europe? A First Evaluation Based on the 2014 EU Labour Force Survey Ad Hoc Module." European Commission Working Paper 1/2016.

Dustmann, Christian, and Francesca Fabbri. 2003. "Language Proficiency and Labour Market Performance of Immigrants in the UK." *Economic Journal* 113 (489): 695–717.

Dustmann, Christian, Francesco Fasani, Tommaso Frattini, Luigi Minale, and Uta Schönberg. 2017. "On the Economics and Politics of Refugee Migration." *Economic Policy* 32 (91): 497–550.

Dustmann, Christian, and Joseph-Simon Görlach. 2015. "Selective Out-Migration and the Estimation of Immigrants' Earnings Profiles." In *Handbook of the Economics of International Migration*, Vol. 1, edited by Barry R. Chiswick and Paul W. Miller, 489–533. Amsterdam, Netherlands: Elsevier.

Edin, Per-Anders, Peter Fredriksson, and Olof Åslund. 2003. "Ethnic Enclaves and the Economic Success of Immigrants—Evidence from a Natural Experiment." *Quarterly Journal of Economics* 118 (1): 329–57.

Edin, Per-Anders, Peter Fredriksson, and Olof Åslund. 2004. "Settlement Policies and the Economic Success of Immigrants." *Journal of Population Economics* 17 (1): 133–55.

European Commission. 2008 and 2014. "European Union Labour Force Survey." Eurostat. https://doi.org/10.2907/LFS1983-2018V.1 (accessed April 8, 2019).

Evans, William N., and Daniel Fitzgerald. 2017. "The Economic and Social Outcomes of Refugees in the United States: Evidence from the ACS." NBER Working Paper 23498.

Fasani, Francesco, Tommaso Frattini, and Luigi Minale. 2018. "(The Struggle for) Refugee Integration into the Labour Market: Evidence from Europe." IZA Discussion Paper 11333.

Fix, Michael, Kate Hooper, and Jie Zong. 2017. "How Are Refugees Faring? Integration at US and State Levels." Washington, DC: Migration Policy Institute.

García, Wendy. 2019. "Why I'm Fleeing Honduras to Seek Asylum in the US." *The Guardian*, July 29. https://www.theguardian.com/environment/2019/jul/29/honduran-asylum-seeker-dam-protester.

German Institute for Economic Research. 1984–2017. "Socio-Economic Panel-Core Version 34." Socio-Economic Panel. doi:10.5684/soep.v34 (accessed May 8, 2019).

German National Academy of Sciences Leopoldina. 2018. *Traumatised Refugees—Immediate Response Required*. Halle, Germany: German National Academy of Sciences Leopoldina.

Giacco, Domenico, Neelam Laxhman, and Stefan Priebe. 2018. "Prevalence of and Risk Factors for Mental Disorders in Refugees." *Seminars in Cell and Developmental Biology* 77: 144–52.

Giuntella, Osea, Zovanga Kone, Isabel Ruiz, and Carlos Vargas-Silva. 2018. "Reason for Immigration and Immigrants' Health." *Public Health* 158: 102–9.

Goebel, Jan, Markus M. Grabka, Stefan Liebig, Martin Kroh, David Richter, Carsten Schröder, and Jürgen Schupp. 2019. "The German Socio-Economic Panel Study (SOEP)." *Journal of Economics and Statistics* 239 (2): 345–60.

Hainmueller, Jens, Dominik Hangartner, and Duncan Lawrence. 2016. "When Lives Are Put on Hold: Lengthy Asylum Processes Decrease Employment Among Refugees." *Science Advances* 2 (8): E1600432.

Harder, Niklas, Lucila Figueroa, Rachel M. Gillum, Dominik Hangartner, David D. Laitin, and Jens Hainmueller. 2018. "Multidimensional Measure of Immigrant Integration." *Proceedings of the National Academy of Sciences* 115 (45): 11483–488.

Hatton, Timothy J. 2016. "Refugees, Asylum Seekers, and Policy in OECD Countries." *American Economic Review* 106 (5): 441–5.

Hatton, Timothy J. 2017. "Refugees and Asylum Seekers, the Crisis in Europe and the Future of Policy." *Economic Policy* 32 (91): 447–96.

Hatton, Timothy J. Forthcoming. "Asylum Migration to the Developed World: Persecution, Incentives and Policy." *Journal of Economic Perspectives*. https://doi.org/10.1257/jep20191107.

Home Office, UK Border Agency: Analysis, Research and Knowledge Management. 2010. "Survey of New Refugees, 2005–2009." http://doi.org/10.5255/UKDA-SN-6556-1.

Horyniak, Danielle, Jason S. Melo, Risa M. Farrell, Victoria D. Ojeda, and Steffanie A. Strathdee. 2016. "Epidemiology of Substance Use among Forced Migrants: A Global Systematic Review." *Public Library of Science One* 11 (7): E0159134.

Hvidtfeldt, Camilla, Jørgen Holm Petersen, and Marie Norredam. Forthcoming. "Prolonged Periods of Waiting for an Asylum Decision and the Risk of Psychiatric Diagnoses: A 22-Year Longitudinal Cohort Study from Denmark." *International Journal of Epidemiology*.

Hvidtfeldt, Camilla, Marie Louise Schultz-Nielsen, Erdal Tekin, and Mogens Fosgerau. 2018. "An Estimate of the Effect of Waiting Time in the Danish Asylum System on Post-resettlement Employment among Refugees: Separating the Pure Delay Effect from the Effects of the Conditions under Which Refugees are Waiting." *Public Library of Science One* 13 (11): E0206737.

Kaltenbach, Elisa, Maggie Schauer, Katharin Hermenau, Thomas Elbert, and Inga Schalinski. 2018. "Course of Mental Health in Refugees—A One Year Panel Survey." *Frontiers in Psychiatry* 9.

Lundborg, Per. 2013. "Refugees' Employment Integration in Sweden: Cultural Distance and Labor Market Performance." *Review of International Economics* 21 (2): 219–32.

Marbach, Moritz, Jens Hainmueller, and Dominik Hangartner. 2018. "The Long-Term Impact of Employment Bans on the Economic Integration of Refugees." *Science Advances* 4 (9): EAAP9519

Mata, Fernando, and Ravi Pendakur. 2017. "Of Intake and Outcomes: Wage Trajectories of Immigrant Classes in Canada." *Journal of International Migration and Integration* 18 (3): 829–44.

Office of Immigration Statistics. 2001–2017. "Yearbook of Immigration Statistics." https://www.dhs.gov/immigration-statistics/yearbook (March 31, 2019).

Office for National Statistics, Social Survey Division, Northern Ireland Statistics and Research Agency, and Central Survey Unit. 2008. "Labour Force Survey Ad Hoc Module Eurostat Dataset, 2008." UK Data Service. http://doi.org/10.5255/UKDA-SN-6770-1.

Phillimore, Jenny. 2011. "Refugees, Acculturation Strategies, Stress and Integration." *Journal of Social Policy* 40 (3): 575–93.

Porter, Matthew, and Nick Haslam. 2005. "Predisplacement and Postdisplacement Factors Associated

with Mental Health of Refugees and Internally Displaced Persons: A Meta-analysis." *Journal of the American Medical Association* 294 (5): 602–12.

Priebe, Stefan, Domenico Giacco, and Rawda El-Nagib. 2016. "Health Evidence Network Synthesis Report 47: Public Health Aspects of Mental Health among Migrants and Refugees: A Review of the Evidence on Mental Health Care for Refugees, Asylum Seekers and Irregular Migrants in the WHO European Region." Copenhagen, Denmark: World Health Organization Regional Office for Europe.

Refugee Action. 2017. "Adam." October 21. https://www.refugee-action.org.uk/adam/.

Ruggles, S., Flood, S., Goeken, G., Grover, J., Meyer, E., Pacas, J., and Sobek, M. 2019. "IPUMS USA: Version 9.0." IPUMS. https://doi.org/10.18128/D010.V9.0.

Ruiz, Isabel, and Carlos Vargas-Silva. 2017. "Are Refugees' Labour Market Outcomes Different from Those of Other Migrants? Evidence from the United Kingdom in the 2005–2007 Period." *Population, Space and Place* 23 (6): E2049.

Ruiz, Isabel, and Carlos Vargas-Silva. 2018. "Differences in Labour Market Outcomes between Natives, Refugees and Other Migrants in the UK." *Journal of Economic Geography* 18 (4): 855–85.

Sarvimäki, Matti. 2017. "Labor Market Integration of Refugees in Finland." *Nordic Economic Policy Review* 2017: 91–114.

Schock, Katrin, Maroa Böttche, Rita Rosner, Mechthild Wenk-Ansohn, and Christine Knaevelsrud. 2016. "Impact of New Traumatic or Stressful Life Events on Pre-existing PTSD in Traumatized Refugees: Results of a Longitudinal Study." *European Journal of Psychotraumatology* 7 (1): Article 32106.

Schultz-Nielsen, Marie Louise. 2017. "Labour Market Integration of Refugees in Denmark." *Nordic Economic Policy Review* 2017: 55–90.

UN High Commissioner for Refugees (UNHCR). 2011. "Resettlement Handbook." Geneva, Switzerland: UNHCR.

UN High Commissioner for Refugees (UNHCR). 2019a. "Global Trends: Forced Displacement in 2018." Geneva, Switzerland: UNHCR.

UN High Commissioner for Refugees (UNHCR). 2019b. "Mediterranean Situation." Operational Portal Refugee Situations. https://data2.unhcr.org/en/situations/mediterranean (accessed November 17, 2019).

Watson, Stephen T. 2019. "From Kenyan Refugee Camp to UB Commencement, One Graduate's Story." *Buffalo News*, May 18. https://buffalonews.com/2019/05/18/from-kenyan-refugee-camp-to-ub-commencement-one-grads-story/.

Zwysen, Wouter. 2019. "Different Patterns of Labor Market Integration by Migration Motivation in Europe: The Role of Host Country Human Capital." *International Migration Review* 53 (1): 59–89.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 122–144

Does Household Electrification Supercharge Economic Development?

Kenneth Lee, Edward Miguel, and Catherine Wolfram

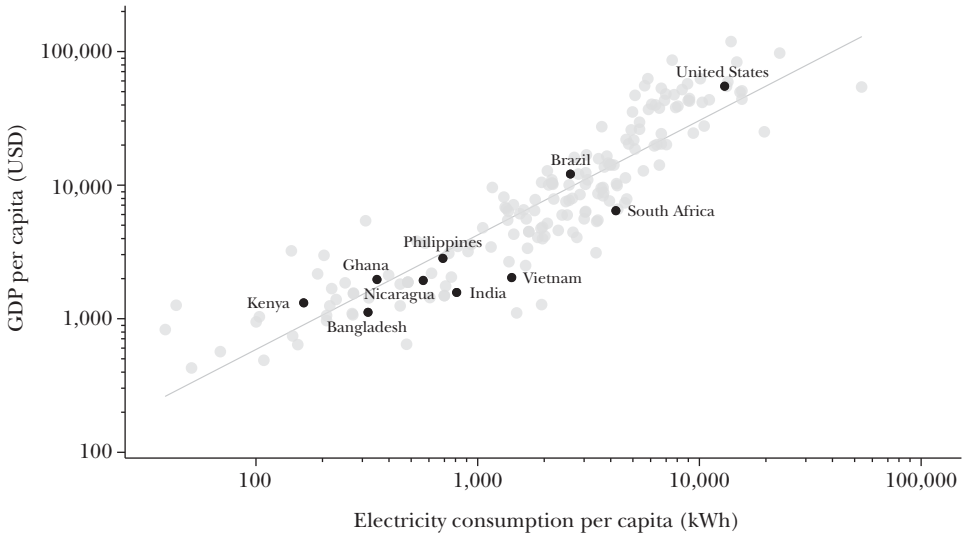
While launching the Sustainable Energy for All program to promote rural electrification in 2011, then-United Nations Secretary General Ban Ki-moon described energy as "the golden thread that connects economic growth, increased social equity, and an environment that allows the world to thrive" (SEFA 2012). Reinforcing this perspective is the strong, positive cross-country correlation between electricity consumption and GDP per capita documented in the macroeconomic literature (for example, Burke, Stern, and Bruns 2018), which we present in Figure 1. Today, nearly a billion people still live without access to electricity (IEA 2018). Thus, access to energy has reemerged as a key priority for policymakers and donors in low-income countries. Electrification could allow poor households to have easy access to lighting for evening chores or studying and power for phone charging and possibly for a range of new small business activities, both on and off the farm.

The idea of a government-subsidized mass electrification program can be traced back to the historical "big push" development efforts of the previous century. In the United States, initiatives like the Tennessee Valley Authority and the Rural

■ *Kenneth Lee is Executive Director of the Energy Policy Institute at the University of Chicago (EPIC India), New Delhi, India. Edward Miguel is Oxfam Professor of Environmental and Resource Economics, University of California, Berkeley, California. Catherine Wolfram is Cora Jane Flood Professor of Business Administration, Haas School of Business, University of California, Berkeley, California. Miguel and Wolfram are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are kennethlee@uchicago.edu, emiguel@berkeley.edu, and cwolfram@berkeley.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.122.

Figure 1
The Positive Correlation between Electricity Consumption and GDP per Capita



Source: 2014 data obtained from the World Bank DataBank.
Note: Both variables are presented on a logarithmic scale. GDP per capita data are in current US dollars.

Electrification Administration, both of which were launched in the 1930s, dramatically expanded electricity generation capacity and rural electrification rates across the American South and other regions. Recent research finds that these programs generated meaningful long-run economic benefits (Kline and Moretti 2014; Kitchens and Fishback 2015; Lewis and Severnini, forthcoming).

Nearly a century later, substantial investments are still being made to expand energy access around the world. The focus of some of the influential development policies that are in place today—like Sustainable Development Goal 7 from the United Nations, which targets universal access to energy by 2030, and the US Power Africa initiative, which aims to add 60 million new electricity connections across Africa—is largely placed on increasing household electrification rates. But the evidence on how much, and in what ways, modern-day residential electrification alone contributes to economic development is not always clear and is sometimes in conflict.

In this paper, we discuss what we can learn from the past decade of microeconomic research on the impacts of household electrification, with the goal of highlighting how future initiatives can be better designed. We begin with an overview of how household electrification has traditionally been captured in official statistics and then turn to some of the historical electrification programs from around the world, paying special attention to those that are most closely related to the settings that have been studied over the past decade or so.

Broadly, the earlier research from this period suggests that access to electricity is a driver of economic development. At the regional level, electrification appears to increase manufacturing output (Rud 2012) and agricultural and manufacturing employment (Kline and Moretti 2014), along with the UN Human Development Index and average housing values (Lipscomb, Mobarak, and Barham 2013). At the household level, which is the focus of this paper, electrification leads to improvements in summary measures of well-being, such as income, expenditure, and consumption (IEG 2008; Khandker, Barnes, and Samad 2012; Van de Walle et al. 2017; Chakravorty, Emerick, and Ravago 2016). The primary mechanisms through which electrification affects development outcomes include: increases in labor supply, particularly for women (Dinkelman 2011, Grogan and Sadanand 2013); higher schooling attainment for children (Khandker et al. 2014, Akpandjar and Kitchens 2017); and better respiratory health (Barron and Torero 2017); among others.[1]

However, a number of these studies rely on relatively strong and untested econometric assumptions, making it a challenge to disentangle the causal effects of electrification on development outcomes from other factors that may also be changing with electrification rates. There may also be lingering reverse causality issues, since economic growth—current or anticipated—may in turn drive greater electricity consumption. More recent studies exploiting experimental or quasi-experimental designs find far less pronounced impacts of electrification on both economic and noneconomic outcomes, most of which are statistically indistinguishable from zero, at least in the medium run (Burlig and Preonas 2016; Lee, Miguel, and Wolfram, forthcoming).

Here, we do not seek to conduct a comprehensive literature review, given that there is already excellent work along these lines: for examples, see Bayer et al. (2019) for a systematic review; Van de Walle et al. (2017) for a general literature review; Morrissey (2018) for a discussion on productive uses of electric power; Peters and Sievert (2016) for a discussion of the studies using African data; and Bernard (2012) for historical context on electrification initiatives in Sub-Saharan Africa. Instead, we attempt to fill a gap in previous reviews by discussing why the existing set of studies might reach such different conclusions, focusing on differences in econometric methods, the types of electrification interventions studied, the potential for spillovers, and differences in regions and populations. To demonstrate how impacts can vary across subgroups of the same population, we build upon the randomized controlled trial design in Lee, Miguel, and Wolfram (forthcoming) to estimate the heterogeneous treatment effects of household grid connections in rural Kenya. We find suggestive evidence that greater gains from electrification are

[1] A related literature addresses how low- and middle-income country firms respond to electricity shortages (the intensive margin) instead of the presence or absence of electricity (the extensive margin). Generally, firms invest in backup generators as a substitute for grid electricity (Steinbuks and Foster 2010), which can limit their overall productivity losses (Allcott, Collard-Wexler, and O'Connell 2016); outsource, essentially substituting electricity inputs with other types of intermediate inputs (Fisher-Vanden, Mansur, and Wang 2015); or switch to more electricity-efficient technologies (Alam 2013).

likely to be concentrated in certain subgroups of households. In our example, the greater gains from electrification occur in households that are willing to pay more for an electricity connection at baseline.

Our main point is that providing poor households with access to electricity alone is not enough to improve economic and noneconomic outcomes in a meaningful way. The literature documents large gains from electrification in a number of settings, but in many cases, we cannot rule out the possibility that other factors—either correlated with or visibly part of the electrification efforts—are driving economic outcomes. Universal energy access is arguably an important goal for global equity considerations. But large-scale contemporary initiatives to expand residential access to electricity may not produce meaningful economic impacts unless they are combined with complementary programs that will make electrical appliances more accessible, or they are targeted towards regions that already benefit from complementary factors.

Measuring Access to Electricity

How electrification is defined and measured is important because it shapes our views on the nature of energy poverty and the solutions that are required. Access to electricity has historically been characterized as a binary state: that is, households have either been considered "on-grid" or "off-grid." In the World Bank's World Development Indicators database, for example, the only regularly tracked electrification data point is "access to electricity," which is presented as a simple percentage of the population and, crucially, is only recorded for the residential sector.

But electrification is clearly more than a binary variable. The term "off-grid," for example, evokes images of remote, rural households that are too far away to connect to power. In Lee et al. (2016), we demonstrate how, just prior to the recent rapid expansion of the rural electricity grid in Kenya, the majority of households were "under-grid," or close enough to be connected to a low-voltage line at a reasonable cost. This distinction matters because the appropriate policy responses for under-grid communities (which could potentially be connected to the grid) may be different from those for truly off-grid communities, which may require the large-scale expansion of national grid infrastructure or stand-alone minigrid or microgrid systems. Another dimension of access to electricity is the reliability of service, an issue that plagues grid-connected households in many low- and middle-income countries. In Nigeria, the electricity connection rate was nearly 60 percent in 2016, but the reliability of electricity was so poor that most people needed to obtain their power from small, diesel generators (as reported in Onishi 2015).

Efforts are underway to expand the way household electrification is measured. The World Bank's Energy Sector Management Assistance Program (ESMAP), for instance, has introduced a new approach called the Multi-tier Framework, in which the measured level of electrification gradually increases with the capacity, duration, reliability, quality, affordability, legality, and safety of electricity access

(available at https://www.esmap.org/node/55526). But for now, we still lack basic data describing how energy poverty varies across space, both in access and in reliability. Even with an expanded delineation of household access, variation in electricity services for nonresidential customers—including factories, small businesses, schools, health centers, and others—will remain unmeasured. This has been a common limitation across most of the existing literature, which collapses all variation in electricity access into a single indicator. We return to this issue later in this paper when discussing differences in the types of interventions studied.[2]

Electrification Initiatives and Estimates

In Table 1, we summarize some of the historical rural electrification efforts that are closely related to the settings studied in the recent microeconomics literature.[3] For each initiative, we note the national and rural electrification rates and GDP per capita at the start and end of the electrification period.

What immediately stands out is how many of these initiatives differ from one another. For example, consider the wide range of starting income and electrification levels across the various initiatives. In the United States, the Rural Electrification Administration was formed in 1935 when GDP per capita was about $9,644 (in 2017 dollars), roughly eight times higher than the GDP per capita in Kenya and India at the beginning of their own respective initiatives. Based on the difference in average income levels alone, it is plausible that newly electrified households and farms in the 1930s United States would have been much better positioned to acquire complementary inputs to electrification, compared to their more recent counterparts in Kenya and India.

The US Rural Electrification Administration was distinctive for several other reasons as well. First, unlike the more recent initiatives in Kenya and India (in which government programs directly connected households and villages to the grid), it was designed to provide low-interest loans to newly formed agricultural cooperatives that were themselves responsible for connecting farms to the grid and paying back the loans. Second, it was introduced at roughly the same time as a number of other New Deal-era programs—including public works programs and fiscal and monetary reforms. Also, it involved efforts to promote and raise awareness about the productive agricultural applications of electricity—such as cooled milk storage and spray irrigation—as well as domestic applications like electric lighting, heated water, electric stoves, and washing machines (Kitchens and Fishback 2015). There was also an associated financing program to facilitate household purchases of appliances. We raise

[2] In the online Appendix available with this paper at the *Journal of Economic Perspectives* website, we present an example of a new approach to capturing energy poverty across Africa in terms of "missing" night lights, based on the difference between local population density and nighttime brightness, presented in online Appendix Figure 1.

[3] This list includes many large economies, although China is absent. We speculate that the list of countries largely reflects settings in which there is appropriate data for research.

Table 1

Historical Rural Electrification Initiatives

Country	Major initiative	Change over period: Electrification: National (%)	Change over period: Electrification: Rural (%)	Change over period: GDP ($/cap.)	Est. cost ($ bn)
USA 1935–1960	*Rural Electrification Administration (REA)*: Provided low-interest loans to newly formed cooperatives to fund rural electrification as part of the *New Deal*, which included fiscal and monetary reforms, public works projects, and new regulations.	67 to 98	< 10 to 96	9,644 to 19,678	4.0 (between 1935 and 1939)
Brazil 1960–2000	*Eletrobras Power Distribution Projects I, II*: Between 1982 and 1991, Eletrobras I and II strengthened distribution networks, expanded supply, and increased rural access rates from 19 to 49 percent. The period also witnessed public investments across various sectors as well as policies to counter hyperinflation.	n/a to 94	< 10 to 75	2,929 to 6,813	24.4 (between 1982 and 1991)
Bangladesh 1977–present	*Rural Electrification Board (BREB)*: Since the 1970s, BREB targeted universal access and other institutional improvements in rural areas that have also benefited from social mobilization campaigns related to health, education, financial inclusion, and others.	n/a to 75	< 10 to 69	470 to 1,524	4.4 (as of 2016)
India (I) 1982–1999	*Integrated Rural Energy Program (IREP)*: Aimed to increase institutional capabilities to meet domestic energy needs (catered towards agricultural and rural development) as part of the *Minimum Needs Program*, which covered rural water supply, health, housing, roads, and others.	n/a to 60	24 to 71	456 to 834	n/a
Ghana 1989–present	*National Electrification Program (NEP)*: Launched in 1989, NEP targeted universal access by 2020, focusing first on major population centers, while the *Self Help Electrification Program (SHEP)* aimed to connect rural areas within 20 kilometers of an existing transmission line.	23 to 78	n/a to 625	66 to 1,338	625
South Africa 1994–1999	*National Electrification Programme (NEP)*: Targeted 2.5 million new household connections, mainly in disadvantaged and rural areas, and all schools and clinics as part of the newly elected government's *Reconstruction and Development Programme*, which initiated large investments across multiple sectors.	36 to 66	12 to 46	4,390 to 4,559	1.6
Vietnam 2000–2006	*Vietnam Rural Energy Project I*: After the end of the US trade embargo, Vietnam established its state utility and enacted power sector reforms. In 2000, the focus shifted towards remote, unelectrified communes and villages.	86 to 96	70 to 92	926 to 1,306	0.3 (between 2000 and 2007)
Philippines 2004–present	*Expanded Rural Electrification Program*: Targeted electrification of all villages by 2008 and 90 percent of households by 2017, mainly by providing low-cost financing to cooperatives and promoting private sector investments.	73 to < 85	n/a to 74	1,899 to 3,105	n/a
India (II) 2005–present	*Rajiv Gandhi Grameen Vidyutikaran Yojana*: The RGGVY program aimed to enhance electricity access in over 400,000 village and connect more than 23 million households. National road connectivity and social security programs for rural areas were also implemented during this period.	67 to 84	57 to 78	1,084 to 2,193	12.9 (between 2012 and 2022)
Kenya 2007–present	*REA and Last Mile Connectivity Project*: Rural Electrification Authority (REA) focused on connecting rural public facilities (for example, schools, clinics, and markets). The Last Mile Connectivity Project (LMCP), which was first announced in 2015, is targeting universal access for households by 2030.	24 to 56	14 to 39	1,232 to 1,541	> 1.0 (including LMCP)

Note: All GDP figures are in 2017 USD. For ongoing initiatives, end-years report statistics for 2017, the latest available year. See the online Appendix Note 1 for further details and references.

this example to highlight the contextual factors that may have also contributed to the success of the US electrification experience.

How have researchers estimated the impact of electrification on household economic development outcomes across these various episodes? Nearly all existing studies use economic survey data to estimate versions of a regression equation in which the dependent variable is a key outcome of interest like labor supply or schooling years for an observed unit (typically a household or a region) at a certain point in time, and the key explanatory variable is a measure of electrification, which is typically a binary variable indicating whether a household has an electricity connection.

Obvious issues arise if the coefficient on the electrification variable is interpreted as capturing the causal effect of switching from no connection to an electricity connection. The primary challenge is that electrification is likely to be correlated with other factors that jointly determine current and expected levels of the outcomes of interest. For example, consider a setting in which there were no subsidies for electricity connections. The households that are connected to power are probably those with higher incomes, wealth, access to credit, and education, or those who believe they would benefit most from an electricity connection. It would be misguided to conclude that any differences between connected and unconnected households can be attributed to differences in electricity access alone.

Similarly, consider how a government (or electric utility) might plan its rollout of electricity infrastructure. If political concerns are prioritized, electric-grid investments may be targeted towards districts that are favored by a ruling government party, and these same districts could also be in line to benefit from a myriad of other government assistance programs.[4] Here, the electrification variable would capture a broader pattern of government favoritism. Alternatively, they may be targeted towards areas that are predicted to have greater potential for economic growth, perhaps due to the presence of a valuable local commodity or the establishment of a new industry that will attract additional labor, further boosting local economic activity. Clearly, it would be misguided to conclude that extending electrification to areas lacking this potential would generate the same effects.

In these examples, omitted variable bias would lead the analyst to overestimate the causal effect of electricity. Of course, these issues can be addressed using various well-known econometric strategies, including difference-in-differences, instrumental variables, regression discontinuity designs, randomized controlled trials, and other methods. But even amongst studies that use these methods, the past decade of work on this topic has resulted in a wide range of estimated effects.

To illustrate this point, we focus on two important household outcomes of electrification that are prominently studied in the recent microeconomics literature: labor supply and education. Following the seminal work of Dinkelman (2011)

[4]For example, Min and Golden (2014) find evidence that politicians in India may manipulate the supply of electricity (for example, by allowing more theft to occur) to influence the outcomes of upcoming elections.

on South Africa's experience with rural electrification in the 1990s, numerous studies have examined whether electrification affects the allocation of household labor resources. The leading hypothesis is that the availability of electricity inside a home reduces the amount of time required for certain household tasks, and that this primarily frees women to pursue and benefit from external employment opportunities.

In Figure 2, panel A, we present key estimates of the impact of electrification on labor supply, separating by male and female wherever possible. In order to compare different studies on the same scale, each coefficient estimate is expressed as a percentage of the mean of the dependent variable. Along the bottom of the figure, we note the econometric strategy used to address the core identification problem for each estimate. In South Africa, rural electrification led to a large 9 to 9.5 percentage point increase in local female employment on a mean of 7 percent baseline female employment (Dinkelman 2011). Similarly, large positive results are documented in Brazil (Lipscomb, Mobarak, and Barham 2013) and Nicaragua (Grogan and Sadanand 2013), two other studies that use instrumental variable approaches.

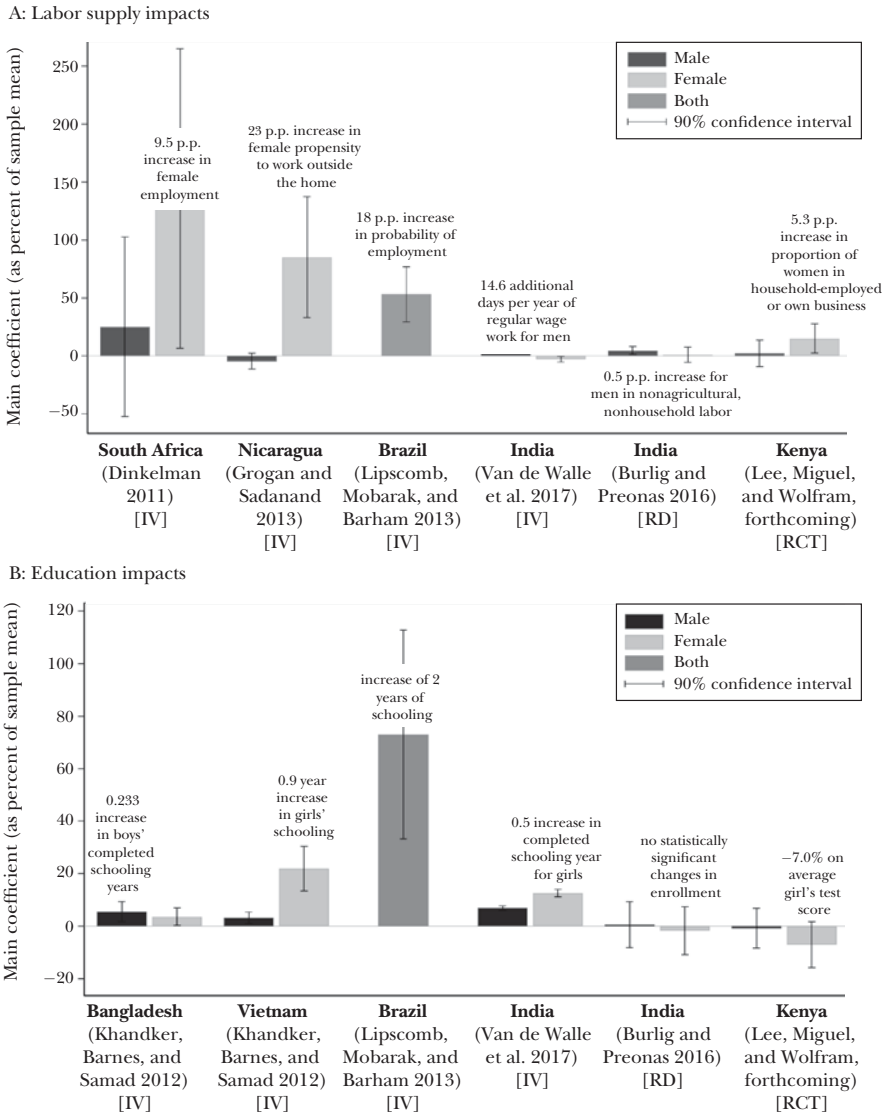
In more recent work, however, the pattern of a large and positive impact on female labor supply seems to disappear. For instance, Van de Walle et al. (2017) find only a small effect in rural India using an instrumental variable approach; Burlig and Preonas (2016) find no economically or statistically significant effect in rural India using a regression discontinuity design; and in Lee, Miguel, and Wolfram (forthcoming), we find only a modest effect for women (and almost no effect for men) in rural Kenya using a randomized controlled trial.

Similarly, in Figure 2, panel B, we present key estimates of the impact on education-related outcomes, again separating for boys and girls wherever possible. In theory, electrification introduces the possibility of electric lighting, which allows children to study for longer hours in the evening, and this may result in improved test scores and higher schooling attainment. Similar to the labor supply findings, the earlier set of studies suggest that electrification has large, positive impacts on education-related outcomes. In Vietnam, Khandker, Barnes, and Samad (2013) use an instrumental variable approach to estimate a 0.9 year increase (21.9 percent) in schooling for girls. But more recent studies in India and Kenya find no statistically significant changes in school enrollment or test scores, using instrumental variable, regression discontinuity, and randomized controlled trial approaches.[5]

How can we make sense of these conflicting results? In the next section, we discuss the role of differences in econometric methods, interventions, levels of measurement, regions, and populations in explaining these patterns.

[5]This pattern is also observed in a comparison of key estimates of the impacts of electrification on income, presented in online Appendix Figure 2.

Figure 2
Key Estimates of the Impacts of Rural Electrification



Source: Author calculations, based on the estimates presented in each of the cited articles.
Note: In this figure, we present key estimates of the impact of electrification on labor supply (panel A) and education (panel B) outcomes. For each study, coefficient estimates have been expressed as a percentage of the mean of the dependent variable. Percentage point units are denoted as p.p.

Making Sense of Divergent Estimates

Different Methods

Electricity grid infrastructure is costly and long-lived, and its planning and construction requires the inputs of multiple stakeholders. Thus, it is rarely randomized and instead is likely to be endogenous to a variety of economic and political factors. Although all of the studies presented in Figure 2 attempt to address selection bias in their own way, each approach relies on a set of assumptions.

Dinkelman (2011), for example, employs an instrumental variables method, utilizing land gradient as an instrument for the wave of rural electrification that followed the end of apartheid in South Africa. A higher land gradient raises the average construction cost of a household connection, and so it is likely to factor into the probability of electrification. In addition, it is not immediately clear why land gradient would be correlated with local employment other than through its effect on construction costs. Thus, it is plausible that using land gradient in an instrumental variable approach can produce unbiased estimates of impacts.[6]

Many of the studies on electrification use an instrumental variable approach in a similar way and attempt to isolate the variation in the electrification variable that can be attributed to a set of exogenous cost considerations. Lipscomb, Mobarak, and Barham (2013), for example, use a time series of hypothetical electricity grids—that simulate how the grid would have evolved had investments been based solely on geographic cost considerations—as an instrument for the actual evolution of the electricity grid in Brazil. Other studies construct instrumental variables based on distances between households (or communities) and the nearest grid infrastructure, assuming that proximity to existing infrastructure is correlated with the cost of grid extension but uncorrelated with current and future economic outcomes (for example, Khandker, Barnes, and Samad 2012; Van de Walle et al. 2017; Chakravorty, Emerick, and Ravago 2016). This approach is feasible and especially appealing considering the growing richness and availability of spatial economic data.

However, it is hard to rule out the possibility that the correlation between the instrument and the dependent variable runs through additional channels beyond electrification. Returning to the case of South Africa, land gradient may have been equally likely to have influenced the cost and placement of post-apartheid roads (or other infrastructure). Roads can reduce transportation time, making it cheaper to visit market centers, improving the conditions for local employment and other economic outcomes. This possibility raises questions about the validity of any geographic cost-based instrument, including in South Africa. During the same post-apartheid period, a large number of public investments were made across multiple

[6] In technical terms, this is the same as saying that the "exclusion restriction" should hold. Note that the instrumental variable method requires that an instrument is *informative* (that is, $E(z_i E_i) \neq 0$, where z_i is the instrument and E_i is the electrification status for household i) and *valid* (that is, $E(z_i \varepsilon_i) = 0$, where ε_i is the error term in the regression described in the previous section). The latter condition is referred to as the "exclusion restriction."

sectors, and as with rural electrification, these investments were also largely targeted towards relatively poor and disadvantaged communities by the newly elected government of President Nelson Mandela. Of course, researchers are well aware of these issues and have made efforts to address them.[7] But in our view, it is difficult to be confident that all of the possible violations of the exclusion restriction have been eliminated. This is especially the case if electrification can interact positively with some unobserved and time-varying factors, as this would result in overestimating the treatment effect.

More recent work has addressed these concerns using alternative econometric strategies. Burlig and Preonas (2016), for example, utilize a regression discontinuity design method, exploiting a population-based eligibility cutoff in India's Rajiv Gandhi Grameen Vidyutikaran Yojana (RGGVY) scheme, a massive national rural electrification program launched in 2005. When certain types of assignment rules (in this case, a cutoff based on village population) are followed, the regression discontinuity design method removes selection bias (here, by comparing villages immediately above and below the cutoff). However, these rules are not always cleanly implemented in low-income countries, forcing researchers to utilize "fuzzy" regression discontinuity design approaches. Burlig and Preonas, however, use satellite images of night lights to show that the RGGVY program did increase electricity availability and consumption, providing supportive evidence that the village population-based cutoff was implemented to a meaningful degree. As noted earlier, they find no evidence of economically or statistically significant impacts on village labor market or educational outcomes.

The obvious hurdle to implementing a randomized controlled trial of electricity grid infrastructure is that researchers find it hard to persuade policymakers to randomize the placement of infrastructure. The Lee, Miguel, and Wolfram (forthcoming) study in rural Kenya, which we revisit later in this paper, is an exception.[8] Like Burlig and Preonas (2016), we find no evidence of meaningful economic, educational, or other impacts among rural households.

Beyond the econometric approach, a common difference between studies that use randomized controlled trials and those that use other methods is the nature of data collection. In an experiment, researchers can design the questions administered through household surveys. As a result, it is possible to collect data on a wider range of outcomes and potential mechanisms than are typically available in the national administrative data that are often used in nonexperimental studies.

[7] Dinkelman (2011) addresses this concern by running a placebo test and other robustness checks. Bensch, Gotz, and Peters (2019) perform alternative placebo tests and show that land gradient is correlated with employment outcomes in nonelectrified areas, suggesting a violation of the exclusion restriction. They provide evidence that land gradient also influenced road placement.

[8] To our knowledge, the only other randomized controlled trials of household electricity connections are: Barron and Torero (2017), which evaluates the impacts of grid connections in El Salvador on indoor air pollution and respiratory outcomes, and Bernard and Torero (2015), the first study that varies grid connections experimentally, which tests for the presence of social interaction effects in driving take-up decisions in Ethiopia but does not evaluate economic outcomes.

In our experiment in Kenya, for example, we collected a variety of information on energy-related outcomes, such as how much each household recently spent on electricity versus kerosene, the variety of electrical appliances owned and desired, the frequency of blackouts recently experienced, and so on. The majority of the studies summarized in Figure 2 are unable to utilize these types of data. The flip side is that administrative data are often more representative and have many more observations, which offers benefits in terms of external validity and statistical power.

Different Interventions and Potential for Spillovers

Another factor contributing to the lack of consensus across studies is that the underlying intervention captured by the electrification variable is not always the same. For instance, the quality of an electricity connection probably varies across programs in terms of the reliability and capacity of power supplied, both of which influence the potential things one can do with electricity.

The design or scale of an electrification program can also result in local spillovers that are not easily measurable using household data. Many historical initiatives to expand electricity access were not only large in scale but also included investments in generation capacity, transmission lines, and other forms of public infrastructure. In Brazil, for example, Lipscomb, Mobarak, and Barham (2013) study the impacts of an electrification effort that entailed a massive upgrade to the nation's energy system. Over the second half of the twentieth century, Brazil witnessed a dramatic expansion in electricity access—the transmission network expanded from 2,359 kilometers in 1950 to 167,443 kilometers in 2000—and substantial investments were also made to increase generation capacity. Much of this progress is owed to Eletrobras, the national electricity utility first established in 1961, which spearheaded the financing and coordination of electricity projects across the country.

If an electrification program is likely to have generated local spillovers, the unit of measurement is important. Studies that measure impacts at the household level will not capture these spillovers to the same extent as studies that observe outcomes at the regional level. In the example of Brazil, Lipscomb, Mobarak, and Barham (2013) measure impacts over a long time frame and at the county level, so any potential within-county economic spillovers are usefully captured in the estimates. Of course, the gains in the Brazil program and related cases flow from not just electrifying households but also schools, health clinics, and local enterprises, making these estimates less comparable to some recent electrification efforts that have targeted households. These features make the Brazil results more comparable to the historical US studies.

Different Regions and Populations

A simple point, but worth emphasizing, is that the effects of household grid connections depend on what individuals are able to do with electricity. As a result, impact estimates may differ across local regions or even across individuals within the same society. Across regions, differences may arise due to the presence or absence of local infrastructure and amenities. For instance, electrification may yield greater

impacts in regions with better access to roads and linkages to neighboring commercial centers, as noted earlier. Impacts may also be greater in areas with existing industries that can benefit from cheaper sources of power or in regions that are experiencing rising income levels due to external factors, like commodity price shocks. Fetter and Usmani (2019), for example, revisit the regression discontinuity design setting studied in Burlig and Preonas (2016) and demonstrate that the impact of India's RGGVY program on nonagricultural employment was higher in villages that simultaneously benefited from a boom in the price of a local commodity (guar). At the same time, Kline and Moretti (2014) find that the magnitude of benefits from the Tennessee Valley Authority program was the same across counties, regardless of whether they were more agricultural or featured any manufacturing at baseline, suggesting that further research into the nature of heterogeneous electrification treatment effects would be useful.

Across individuals within the same society, effects may differ due to variation in individual income levels or access to credit. Wealthier households, by virtue of their ability to purchase more electrical appliances, are likely to be better positioned to benefit from access to electricity. Khandker et al. (2014) is one of the early studies to use an econometric approach to address this question. Using a cross section of household survey data in India, they estimate a quantile regression of overall household income and expenditure on household electrification, addressing the endogeneity of their electrification variable with an instrumental variable strategy. Their analysis—which relies on the arguably strong assumption that the community electrification rate is a valid instrument for household electrification—suggests that households in the highest quintile of income experience nearly double the expenditure impacts as households in the middle quintile. In the following section, we explore this possibility further, exploiting experimental variation from our randomized controlled trial research design in rural Kenya.

The studies discussed in this section offer important contributions to the literature on the impacts of electricity infrastructure, and each utilizes a creative and novel way to address the endogeneity of the electrification variable. But in our view, some skepticism of instrumental variables strategies based on geographic variation is warranted. In addition, it is important to consider the type of electrification intervention, as well as the other amenities that are being made available either through the electrification program or exogenously, as these factors could influence the magnitude of estimated impacts.

New Experimental Evidence on Heterogeneous Treatment Effects

Many existing analyses of heterogeneity rely on the inclusion of interaction terms in the regression specifications between a household's electrification status and observable covariates at baseline, like income, assets, and so on. Here, we build on the randomized controlled trial design in Lee, Miguel, and Wolfram (forthcoming) to show what we can learn from an alternative approach to analyzing

heterogeneity that compares households based on how much they are willing to pay for an electricity connection, a household characteristic that is rarely if ever captured in observational datasets.

In our experiment, we provided randomly selected clusters of households in rural Kenya with an opportunity to connect to the grid at a subsidized price. In order to estimate a demand curve for grid connections, we randomly assigned the connection price across treatment communities. Specifically, one-third of the 75 treatment communities were offered a 29 percent subsidy to connect to the grid (that is, the effective price of a grid connection was reduced from the prevailing official price of $398 to $284); one-third were offered a 57 percent subsidy (the effective price was $171); and one-third were offered a full subsidy (the effective price was $0). Take-up varied dramatically across treatment arms: 95 percent of households accepted a fully subsidized connection; 28 percent took up at a 57 percent discount; and just 14 percent of households paid for a connection at a 29 percent discount, while even fewer control (unsubsidized) households connected to the grid over the study period.

Exogenous variation in electrification status, created by the randomized price offers, generated unbiased estimates of the impacts of electrification. Roughly 16 to 32 months after installation of a home grid connection, the average household showed little evidence of any meaningful economic or noneconomic gains across a wide range of outcomes. Results are similar for the simpler comparison between the control group (in which almost no households were connected) and the full subsidy treatment group (in which nearly all households were connected).

How do these impacts vary across different population groups in this setting? Drawing on standard properties of "local average treatment effects" (related to the discussion in Kowalski 2016), we can separately estimate impacts for different types of households. Specifically, households in our experiment can be allocated into the following complier subgroups: (1) "never takers," meaning households that would not even accept a free connection; (2) "adopters of electricity only when the price is low," meaning households that are willing to accept a connection when the price is $0 (one of the randomly assigned prices) and potentially up to $171; (3) "adopters of electricity when the price is high," meaning households that are willing to accept an electricity connection when the price is between $171 and $284; and (4) "always takers," meaning households that would pay more than $284. In the remainder of this section, we assess whether the subgroup of households that are willing to pay more for electricity—which may be correlated with wealth, access to credit, or to other unobserved dimensions of ability, ambition, or opportunity—end up benefiting more from an electricity connection than others.

A first step towards deriving treatment effects for different complier groups is to estimate their sample shares. It has long been understood that the average treatment effects can be represented as the weighted average of multiple marginal treatment effects that may differ across subgroups (Heckman and Vytlacil 1999, 2001). In our sample, 67 percent of households are "adopters *only* when the price

is low," and 22 percent are "adopters when the price is high." The small shares of remaining households are either "never takers" or "always takers."[9]

The next step is to estimate separate local average treatment effects for each complier subgroup on a range of household outcomes, including among others: monthly electricity spending, the number of appliance types owned (including mobile phones, radios, televisions, and others), monthly spending on kerosene, the share of household members that are employed or own their own businesses, household asset value, and a measure of recent health symptoms experienced by the household respondent. For the "adopters when the price is high" group, we can obtain these estimates from a two-stage least squares regression in which we drop the high- and low-subsidy treatment arms and regress the various outcomes on an indicator for whether a household has an electricity connection, instrumented with an indicator for whether the household was offered a medium subsidy.[10] For the "adopters when the price is low," we can use the subgroup sample shares and back out the local average treatment effect by invoking the formula for weighted averages. For example, the local average treatment effect for compliers in the $0 treatment group is simply the weighted average of the local average treatment effects for the two complier groups of interest.[11]

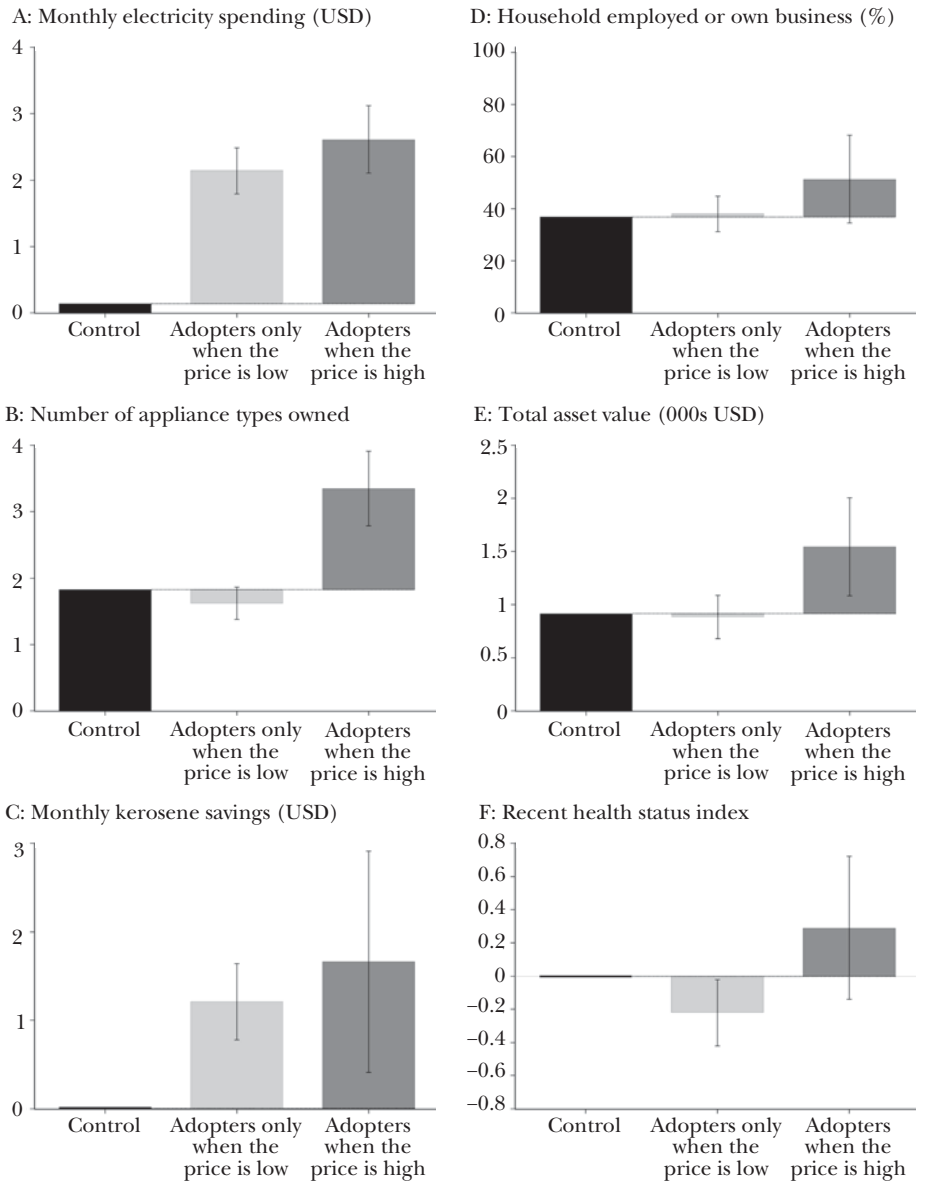
We illustrate the results of this approach in Figure 3, where we compare local average treatment effects for "adopters only when the price is low" against those for "adopters when the price is high" across a key set of outcomes. Overall, "adopters when the price is high" appear to do far more with an electricity connection compared to their counterparts; the figure also contrasts these treatment effects with the mean characteristic in the control (unsubsidized) group. "Adopters when the price is high" spend more on electricity; experience greater savings on kerosene; and acquire a greater variety of appliances, such as mobile phones and televisions. The large difference in the number of appliance types owned across the two complier subgroups—a significant 83 percent for the increase for the "adopters when the price is high" versus a (nonsignificant) 11 percent decrease for those who connect only when it is free—is statistically significant at the 1 percent level. Similarly, "adopters when the price is high" also appear to enjoy more pronounced economic and noneconomic impacts: they are more likely to become employed or own a business, more likely to experience an increase in total asset value, and

[9] Because we randomized price offers across communities, we need only the standard assumption of monotonicity to uncover unbiased estimates of these sample shares (Imbens and Angrist 1994). In online Appendix Note 2, we offer a formal description of our econometric approach to estimating heterogeneous treatment effects across complier subgroups in this setting.

[10] We can do this because compliers in the medium-subsidy group (in which the electricity connection price is $171) include both compliers at $171 as well as compliers at $284 by the monotonicity assumption.

[11] Detailed regression results are available in the online Appendix. In online Appendix Table 1, we report mean values in the control group (column 1), the local average treatment effects for each of the two complier subgroups (columns 2 and 3), and the *p*-value of the difference between the local average treatment effects for each outcome (column 4). Note that we include the same set of variables presented in Table 3 in Lee, Miguel, and Wolfram (forthcoming) to facilitate comparison to the full sample results.

Figure 3
Comparison of Local Average Treatment Effects between Different Complier Groups



Source: Author calculations, based on survey data collected from 2,217 households in western Kenya in 2016.

Note: In this figure, we show how the impacts of household electrification may vary across subgroups of the same population of rural Kenyan households. In Panel C, monthly kerosene savings are presented as relative to the control group mean of $2.81. In Panel F, a positive value reflects a desirable outcome. See online Appendix Table 1 and associated discussion for additional outcomes and details.

more likely to report better health outcomes (note that higher values on the recent health status index correspond to a lower number of recent symptoms reported). In additional results (not shown in Figure 3), we do not find that any subgroup experiences gains in student test scores.

Due to limitations in our sample size—a result of the limited number of households who choose to connect when it is not free—these results should be treated only as suggestive. Many of the estimated local average treatment effects are only marginally significant at traditional confidence levels, and we cannot reject equality of effects across the two complier subgroups in most cases. Yet the pattern of impacts in Figure 3 tells a remarkably consistent story indicating that those who are willing to pay more for an electricity connection are poised to benefit far more than those who only connect when it is free.

Naturally, our approach to estimating heterogeneous treatment effects leads to the question of how households in these complier subgroups are different from one another. Is it possible to identify households that will benefit the most from electrification using a standard set of observable characteristics? We use baseline household survey data collected in 2014 (note that all of these households were unconnected at baseline) to summarize the key differences between the groups.[12] Broadly speaking, "adopters when the price is high" appear to be wealthier and better-off in multiple ways: household heads in this group are more likely to have attended secondary school (21.0 percent versus 9.5 percent), report far higher monthly earnings ($24.39 versus $11.55), and hold a bank account (32.4 percent versus 14.7 percent), with this last difference statistically significant at the 1 percent level. "Adopters when the price is high" also have significantly higher asset ownership. In contrast, several other household characteristics that would seem to be less obviously correlated with wealth, including respondent age and gender as well as the household's distance to the nearest electricity distribution transformer, appear roughly similar across the two groups.

In our example, households that are willing to pay more for an electricity connection also appear to be observably richer and more educated at baseline. However, we cannot rule out the possibility that unobservables—like individual initiative, ambition or "spunk," or other oft-cited unobservables in wage equations—may be correlated with both household wealth, for example, and the ability to make the most of an electricity connection. This possibility suggests that this complier approach to studying heterogeneity, which is possible due to the experimental nature of this study, can be valuable in shedding additional light on how treatment effects vary across individuals. In online Appendix Table A4, we report the results of a regression in which the treatment (household electrification) is interacted with an index of social and economic status at baseline, based on commonly observed measures (like education, income, and others). This approach does not seem to

[12] For a table of results, see online Appendix Table 2.

predict larger effects for households in the top quartile of social and economic status at baseline, in contrast to the approach that compares the two complier groups.

Our main point is that the impacts of electrification can vary substantially across different types of individuals, even within a relatively homogenous sample of poor rural households in neighboring villages, in ways that are difficult to capture with commonly measured household observable characteristics.

Focusing on the Grid

We have largely focused on lessons from the past decade of research on the impacts of residential grid electrification, a growing area of investigation. In addition, the question of how governments can expand electricity access to maximize impact holds a great deal of policy-relevance today. Across Sub-Saharan Africa, where roughly 600 million people are still without power, billions of dollars are being allocated towards expanding residential access to the grid. In Kenya alone, roughly $364 million was committed to the Last Mile Connectivity Project (LMCP) in 2015, in a project that promised to connect four million under-grid households to power (as reported in *Business Daily Africa* (2015) at the official launch of the LMCP in May 2015).

But the grid is just one way to expand electricity access. Since the turn of the current century, countless entrepreneurs, donors, and policymakers have argued that decentralized, renewable energy technologies could allow off-grid households across the developing world to "leapfrog" the conventional grid, similar to how the introduction of mobile phones allowed populations to leapfrog the landline. Indeed, the home solar sector—a term we use to collectively refer to solar lanterns and solar home systems—has seen its estimated penetration rise rapidly across Sub-Saharan Africa. Increasing appliance efficiencies and reductions in the cost of photovoltaics (in addition to improvements in batteries) are some of the factors that may have contributed to this growth (Alstone, Gershenson, and Kammen 2015).

Solar lanterns offer just enough power to meet the basic standard of electrification in the World Bank's Multi-tier Framework, mentioned above. Grid connections can meet far higher standards, depending on their reliability. Increasingly, home solar companies are integrating pay-as-you-go technologies directly into their products, directly addressing the credit constraints that may limit take-up of new technologies in poor settings; in practice, these solar home systems are offered on credit and are remotely disabled if payments are not made on time. Pay-as-you-go has transformed the way these products are marketed, financed, and distributed. In some countries, like Uganda, pay-as-you-go is even allowing consumers to offer their home solar products as collateral for new types of loans (Gertler, Green, and Wolfram 2019).

Separate randomized controlled trials have measured the impacts of home solar access on child study times, finding mixed results: home solar appears to increase study times, but decrease test scores in Uganda (Furukawa 2014); not

increase study times in Kenya (Rom, Günther, and Harrison 2017); and increase study times but only for boys in Rwanda (Grimm et al. 2017). These results highlight the lack of consensus about the educational benefits of home solar. That said, in countless rural households across the world, the increasing adoption of these products should, at the very least, reduce the usage of kerosene and dry cell batteries for lighting, resulting in some benefits to health and the environment.

Microgrids have also generated substantial interest, especially for geographically remote communities that are prohibitively expensive to connect to a national grid. Microgrids are typically defined as small networks of users connected to a centralized and stand-alone source of electricity generation and storage. They are capable of providing longer hours and higher capacities than home solar and can also be powered with clean energy sources like solar, wind, and hydro. Technically, it is possible to integrate them into expanding national grids over the long run, but it is too early to tell how widely this will happen in practice.

Recent research on the demand for microgrid connections has not been wholly positive, at times due to external factors. In Rajasthan, India, for example, Fowlie et al. (2019) document how demand for connections to privately operated solar microgrids is very low, largely due to a perception that the government would soon be subsidizing connections to the central grid. Relatedly, in Bihar, India, Burgess et al. (2019) find that demand for connections to privately operated solar microgrids is strongly influenced by the availability and quality of the central grid. At the same time, a number of private operators have built microgrids in Kenya that are operational and generating revenue, suggesting that demand is positive in some settings.

In addition to expectations about the arrival of the grid, fundamental consumer preferences can also limit the take-up of alternative energy. In Kenya, we document descriptive evidence at baseline, suggesting that home solar does not satisfy a wide range of household energy needs, based on a survey of appliance ownership and aspirations (Lee, Miguel, and Wolfram 2016). Relative to households that primarily use kerosene, home solar users benefit from basic energy applications, including lighting, mobile phone charging, and, for some systems, television. However, once they have access to these basic end uses, the appliances they aspire to own next (for example, irons) require higher wattages that cannot be supported by most home solar systems, at least based on current technologies.

Discussion

Over the past decade, studies on the impacts of residential electrification on the well-being of households in low- and middle-income countries have generated conflicting results. While some studies estimate very large effects on household labor supply, for instance, others rule out point estimates that are even a quarter as large. We explore how differences in methods, interventions, and/or populations may help reconcile these disparate results.

Our main conclusion, based in part on our own recent research, is that the provision of home electrification alone is not enough to improve economic outcomes substantially for the world's poorest citizens. This perspective stands in contrast to the findings in earlier analyses in the literature, which explore electrification impacts in middle-income countries, like South Africa and Brazil. Although retrospective analyses of electrification in the United States in the 1930s point to very large impacts, these initiatives were introduced at a time when GDP per capita (in current dollars) was roughly eight times as large as comparable measures in contemporary Kenya and India. Also, in some cases, the early US initiatives brought electricity to many sectors of the economy, including manufacturing facilities. Reconciling these cross-study differences presents its own identification challenge, as it is hard to know whether these differences are due to the choice of the econometric method, the extent of the electrification initiative, or to relative differences in starting incomes. With that said, our overall position is that the impacts of residential electrification may crucially depend on the extent to which households are positioned to take actions and/or make the complementary investments that will ultimately allow them to make the most out of an electricity connection.

Consistent with this view that context matters, our own recent work finds that heterogenous effects also exist within local areas. We exploit a feature of a recent experiment in western Kenya that allows us to estimate heterogeneous treatment effects across different complier groups using the same identification strategy. We show that households that were only willing to connect to the grid when it was effectively free experience fewer economic gains than households that were willing to connect when the price was high. This result offers suggestive evidence of substantial heterogeneity in treatment effects, even within a sample of poor rural households that were all equally without electricity at baseline.

The question of how the impacts of electrification may vary across countries, or regions within a country, is likely to be of keen policy interest. We see expanding evidence in this area as an important task for future research. The degree of heterogeneity in treatment effects could naturally be much larger across rural and urban areas in the same country or across countries with different income levels. On the one hand, understanding which households and areas are most likely to benefit from grid connections can help policymakers better target grid investments. On the other hand, if wealthier households are more likely to utilize and benefit from access to electricity—due to their ability to make complementary investments or exploit new business opportunities opened up by access to power—expansion of the rural grid infrastructure could exacerbate economic inequality in rural areas of low-income countries, an outcome that is seldom discussed in the current policy debate. This would imply a fundamental tension in rural electrification programs between promoting economic growth and exacerbating inequality.

To date, both policymakers and researchers have often focused on the effects of household electrification. For policymakers, this may reflect either a political calculus that those not presently connected to electricity are a potent group of potential supporters, the belief that electricity should be viewed as a basic right

even for the very poorest citizens, or some combination of the two. In our view, the available evidence suggests that the provision of electrification to poor households is unlikely, on its own, to be economically transformative, at least in the short to medium run. As such, a singular policy focus on electrifying poor and mostly rural households may be misguided. Going forward, we believe that studying the long-run impacts of residential electrification, the interactions between electrification and contextual factors, as well as impacts of electricity access for nonresidential consumers—including schools, health centers, and firms—are all likely to be fruitful research directions.

■ *We are grateful to Felipe Vial, Zachary Obstfeld, Nishmeet Singh, Aishwarya Kumar, and Rongmon Deka for excellent research assistance. An earlier version of this paper was funded with support from the UK government as part of the Department for International Development (DFID) supported by the Energy and Economic Growth (EEG) research program based at the Center for Effective Global Action (CEGA) and the Energy Institute at Haas (EI) at the University of California, Berkeley. We thank Robert Fetter, Gordon Hanson, Enrico Moretti, Timothy Taylor, and Heidi Williams for helpful comments.*

References

Akpandjar, George, and Carl Kitchens. 2017. "From Darkness to Light: The Effect of Electrification in Ghana, 2000–2010." *Economic Development and Cultural Change* 66 (1): 31–54.

Alam, Muneeza M. 2013. "Coping with Blackouts: Power Outages and Firm Choices." https://pdfs.semanticscholar.org/ce0c/5338d15eec64f9227ab52b5cd6e53270e29f.pdf?_ga=2.47862821.369739089.1574588108-2135327386.1574588108 (accessed November 24, 2019).

Allcott, Hunt, Allan Collard-Wexler, and Stephen D. O'Connell. 2016. "How Do Electricity Shortages Affect Industry? Evidence from India." *American Economic Review* 106 (3): 587–624.

Alstone, Peter, Dimitry Gershenson, and Daniel M. Kammen. 2015. "Decentralized Energy Systems for Clean Electricity Access." *Nature Climate Change* 5: 305–14.

Barron, Manuel, and Maximo Torero. 2017. "Household Electrification and Indoor Air Pollution." *Journal of Environmental Economics and Management* 86: 81–92.

Bayer, Patrick, Ryan Kennedy, Joonseok Yang, and Johannes Urpelainen. 2019. "The Need for Impact Evaluation in Electricity Access Research." *Energy Policy.*

Bensch, Gunther, Gunnar Gotz, and Jorg Peters. 2019. "Effects of Rural Electrification on Employment: New Evidence from South Africa—Comment." Unpublished.

Bernard, Tanguy. 2012. "Impact Analysis of Rural Electrification Projects in Sub-Saharan Africa." *World Bank Research Observer* 27 (1): 33–51.

Bernard, Tanguy, and Maximo Torero. 2015. "Social Interaction Effects and Connection to Electricity: Experimental Evidence from Rural Ethiopia." *Economic Development and Cultural Change* 63 (3): 459–84.

Burgess, Robin, Michael Greenstone, Nicholas Ryan, and Anant Sudarshan. 2019. "Demand for Electricity in a Poor Economy." http://www.lse.ac.uk/economics/Assets/Documents/personal-pages/

robin-burgess/demand-for-electricity-in-a-poor-economy.pdf (accessed November 24, 2019).

Burke, Paul J., David I. Stern, and Stephan B. Bruns. 2018. "The Impact of Electricity on Economic Development: A Macroeconomic Perspective." *International Review of Environmental and Resource Economics* 12 (1): 85–127.

Burlig, Fiona, and Louis Preonas. 2016. "Out of the Darkness and into the Light? Development Effects of Rural Electrification." Energy Institute at Haas Working Paper 268.

Business Daily Africa. 2015. "Cost of Electricity Connections Reduced to Sh15,000." May 27. https://www.businessdailyafrica.com/news/Cost-of-power-connection-falls-to-Sh15-000/539546-2731850-phsls5/index.html.

Chakravorty, Ujjayant, Kyle Emerick, and Majah-Leah Ravago. 2016. "Lighting Up the Last Mile: The Benefits and Costs of Extending Electricity to the Rural Poor." https://media.rff.org/documents/RFF-DP-16-22-REV.pdf (accessed November 24, 2019).

Dinkelman, Taryn. 2011. "The Effects of Rural Electrification on Employment: New Evidence from South Africa." *American Economic Review* 101 (7): 3078–3108.

Fetter, T. Robert, and Faraz Usmani. 2019. "Fracking, Farmers, and Rural Electrification in India." https://drive.google.com/file/d/1wI9eCenSWzmY_Xm81Q_UU-kRbJkPpY4l/view (accessed November 24, 2019).

Fisher-Vanden, Karen, Erin T. Mansur, and Qiong (Juliana) Wang. 2015. "Electricity Shortages and Firm Productivity: Evidence from China's Industrial Firms." *Journal of Development Economics* 114: 172–88.

Fowlie, Meredith, Yashraj Khaitan, Catherine Wolfram, and Derek Wolfson. 2019. "Solar Microgrids and Remote Energy Access: How Weak Incentives Can Undermine Smart Technology." *Economics of Energy and Environmental Policy* 8 (1): 33–49.

Furukawa, Chishio. 2014. "Do Solar Lamps Help Children Study? Contrary Evidence from a Pilot Study in Uganda." *Journal of Development Studies* 50 (2): 319–41.

Gertler, Paul, Brett Green, and Catherine Wolfram. 2019. "Unlocking Access to Credit via Lockout Technology." Unpublished.

Grimm, Michael, Anicet Munyehirwe, Jörg Peters, and Maximiliane Sievert. 2017. "A First Step up the Energy Ladder? Low Cost Solar Kits and Household's Welfare in Rural Rwanda." *World Bank Economic Review* 31 (3): 631–49.

Grogan, Louise, and Asha Sadanand. 2013. "Rural Electrification and Employment in Poor Countries: Evidence from Nicaragua." *World Development* 43: 252–65.

Heckman, James J., and Edward J. Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences of the United States of America* 96 (8): 4730–34.

Heckman, James J., and Edward Vytlacil. 2001. "Policy-Relevant Treatment Effects." *American Economic Review* 91 (2): 107–11.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Independent Evaluation Group (IEG). 2008. *The Welfare Impact of Rural Electrification: A Re-assessment of the Costs and Benefits.* Washington, DC: World Bank.

International Energy Agency (IEA). 2018. *World Energy Outlook 2018.* Paris: IEA.

Khandker, Shahidur R., Douglas F. Barnes, and Hussain A. Samad. 2012. "The Welfare Impacts of Rural Electrification in Bangladesh." *Energy Journal* 33 (1): 187–206.

Khandker, Shahidur R., Douglas F. Barnes, and Hussain A. Samad. 2013. "Welfare Impacts of Rural Electrification: A Panel Data Analysis from Vietnam." *Economic Development and Cultural Change* 61 (3): 659–92.

Khandker, Shahidur R., Hussain A. Samad, Rubaba Ali, and Douglas F. Barnes. 2014. "Who Benefits Most from Rural Electrification? Evidence in India." *Energy Journal* 35 (2): 75–96.

Kitchens, Carl, and Price Fishback. 2015. "Flip the Switch: The Impact of the Rural Electrification Administration 1935–1940." *Journal of Economic History* 75 (4): 1161–95.

Kline, Patrick, and Enrico Moretti. 2014. "Local Economic Development, Agglomeration Economies, and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority." *Quarterly Journal of Economics* 129 (1): 275–331.

Kowalski, Amanda E. 2016. "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments." https://cowles.yale.edu/sites/default/files/files/conf/2016/Summer/Econometrics/Kowalski-Doing%20More%20When%20You%27re%20Running%20LATE.pdf (accessed November 24, 2019).

Lee, Kenneth, Eric Brewer, Carson Christiano, Francis Meyo, Edward Miguel, Matthew Podolsky, Javier Rosa, and Catherine Wolfram. 2016. "Electrification for 'Under Grid' Households in Rural Kenya." *Development Engineering* 1: 26–35.

Lee, Kenneth, Edward Miguel, and Catherine Wolfram. 2016. "Appliance Ownership and Aspirations among Electric Grid and Home Solar Households in Rural Kenya." *American Economic Review: Papers and Proceedings* 106 (5): 89–94.

Lee, Kenneth, Edward Miguel, and Catherine Wolfram. Forthcoming. "Experimental Evidence on the Economics of Rural Electrification." *Journal of Political Economy.*

Lewis, Joshua, and Edson Severnini. Forthcoming. "Short- and Long-Run Impacts of Rural Electrification: Evidence from the Historical Rollout of the U.S. Power Grid." *Journal of Development Economics.*

Lipscomb, Molly, A. Mushfiq Mobarak, and Tania Barham. 2013. "Development Effects of Electrification: Evidence from the Topographic Placement of Hydropower Plants in Brazil." *American Economic Journal: Applied Economics* 5 (2): 200–231.

Min, Brian, and Miriam Golden. 2014. "Electoral Cycles in Electricity Losses in India." *Energy Policy* 65: 619–25.

Morrissey, James. 2018. *Linking Electrification and Productive Use.* Washington, DC: Oxfam Research Backgrounder.

Onishi, Norimitsu. 2015. "Weak Power Grids in Africa Stunt Economies and Fire Up Tempers." *New York Times,* July 2. https://www.nytimes.com/2015/07/03/world/africa/weak-power-grids-in-africa-stunt-economies-and-fire-up-tempers.html.

Peters, Jörg, and Maximiliane Sievert. 2016. "Impacts of Rural Electrification Revisited—The African Context." *Journal of Development Effectiveness* 8 (3): 327–45.

Rom, Adina, Isabel Günther, and Kat Harrison. 2017. "The Economic Impact of Solar Lighting: Results from a Randomised Field Experiment in Rural Kenya." https://ethz.ch/content/dam/ethz/special-interest/gess/nadel-dam/documents/research/Solar%20Lighting/17.02.24_ETH%20report%20on%20economic%20impact%20of%20solar_summary_FINAL.pdf (accessed November 24, 2019).

Rud, Juan Pablo. 2012. "Electricity Provision and Industrial Development: Evidence from India." *Journal of Development Economics* 97 (2): 352–67.

Secretary-General's High-Level Group on Sustainable Energy for All (SEFA). 2012. *Sustainable Energy for All: A Global Action Agenda.* Washington, DC: SEFA.

Steinbuks, J., and V. Foster. 2010. "When Do Firms Generate? Evidence on In-House Electricity Supply in Africa." *Energy Economics* 32 (3): 505–14.

Van de Walle, Dominique, Martin Ravallion, Vibhuti Mendiratta, and Gayatri Koolwal. 2017. "Long-Term Gains from Electrification in Rural India." *World Bank Economic Review* 31 (2): 385–411.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 145–169

The Consequences of Treating Electricity as a Right

Robin Burgess, Michael Greenstone, Nicholas Ryan, and Anant Sudarshan

High-income countries take electricity for granted: people know the lights will switch on twenty-four hours a day, 365 days a year. In developing countries, nearly a billion people are not connected to the electricity grid, and those who are receive partial and intermittent power supply. We argue that these shortfalls arise as a consequence of treating electricity as a right, rather than as a private good.

By a "right to electricity," we refer to the social norm that all people deserve electricity regardless of payment. This entitlement has driven universal electrification programs around the world for decades and remains salient in developing countries investing in electrification today. In India, Prime Minister Narendra Modi writes, "Everyone has a right to a life of dignity. Traditionally, food and shelter have been seen as the most basic necessities. However, the Modi government has gone beyond this core basket of necessities to include even electricity" (Modi 2019). In Bolivia, the constitution itself guarantees a right to universal electricity access, and former President Evo Morales declared that electricity, among other basic services, should "be recognized in international legislation and in national standards in all

■ *Robin Burgess is Professor of Economics, London School of Economics, London, United Kingdom. Michael Greenstone is Milton Friedman Distinguished Service Professor in Economics, University of Chicago, Chicago, Illinois. Nicholas Ryan is Assistant Professor of Economics, Yale University, New Haven, Connecticut. Anant Sudarshan is South Asia Director of the Energy Policy Institute at the University of Chicago (EPIC), Chicago, Illinois. Their email addresses are r.burgess@lse.ac.uk, mgreenst@uchicago.edu, nicholas.ryan@yale.edu, and anants@uchicago.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.145.

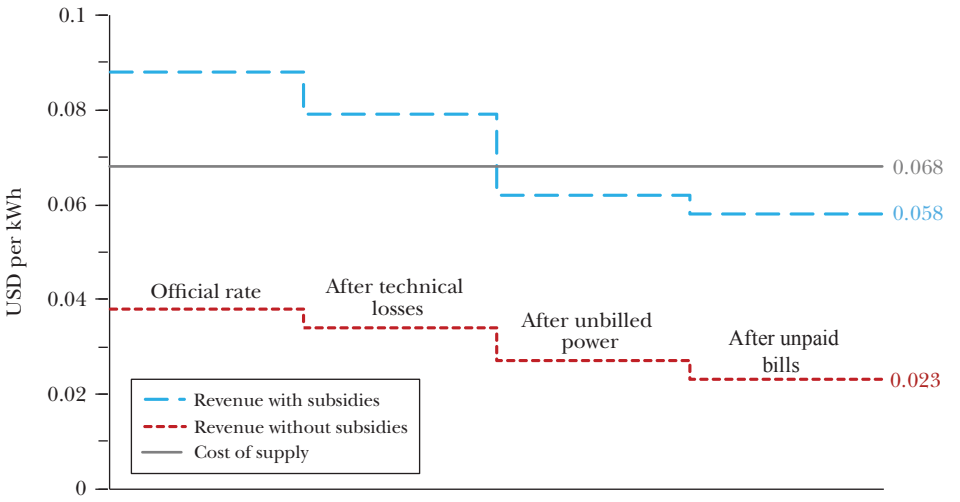
countries as a fundamental human right of the people in all corners of the planet" (Morales 2012). The UN, under Sustainable Development Goal number 7 (SDG7), has set 2030 as the date by which universal access to electricity should be achieved worldwide. Electrification has been described as a necessary step to achieving other goals, including the goals of poverty eradication (SDG1), enhancing education (SDG4), creating economic opportunity (SDG8), and empowering women (SDG5) (SEFA 2012). The push to universal electrification—irrespective of cost—is global and current.[1]

How can treating electricity as a right undermine the aim of universal access to reliable electricity? We argue that there are four steps. In step 1, because electricity is seen as a right, subsidies, theft, and nonpayment are widely tolerated. Bills that do not cover costs, unpaid bills, and illegal grid connections become an accepted part of the system. In step 2, electricity utilities—also known as distribution companies—lose money with each unit of electricity sold and in total lose large sums of money. Though governments provide support, at some point, budget constraints start to bind. In step 3, distribution companies have no option but to ration supply by limiting access and restricting hours of supply. In effect, distribution companies try to sell less of their product. In step 4, power supply is no longer governed by market forces. The link between payment and supply has been severed: those evading payment receive the same quality of supply as those who pay in full. The delinking of payment and supply reinforces the view described in step 1 that electricity is a right. We describe these steps sequentially, but they can be thought of as parts of a low-quality, low-payment equilibrium. This equilibrium, we argue, is what differentiates electricity markets in developed and developing countries.

The consequences for electricity consumers, both rich and poor, are severe. There is only one electricity grid, and it becomes impossible to offer a higher quantity or quality of supply to those consumers who are willing and sometimes even desperate to pay for it. Socially beneficial transactions are therefore prevented from occurring. This interaction of the social norm that electricity is a right and the technological constraint of a common grid for all parties makes it impossible to ration service to person by person, and firm by firm, making the consequences of treating electricity as a right more severe than for other private goods. Though private alternatives to grid electricity exist, like diesel generators and solar panels, these substitutes are inferior to grid electricity in terms of price and load (Burgess et al. 2019). In fact, the only reason these substitutes are competitive at all is that the quality of the service the grid provides is so poor. The lack of close substitutes also differentiates electricity from other goods like water and fuel, which are also

[1] For example, the Indian government is pushing to reach 100 percent household electrification in a context where the unelectrified are rural and poor. The USAID-led Power Africa program aims to add 60 million new electricity connections by 2030, with a focus on infrastructure construction rather than bill payment. Meanwhile, the DFID-led Energy Africa campaign is targeting universal energy access by 2030 both through grid and off-grid sources.

Figure 1
Electricity Losses in Bihar, India



Source: Statistics are from the 2019–2020 BERC Tariff Order and Tariff Schedule.
Note: Bihar-wide losses from June–September 2018 are reported. The solid line is the average power purchase cost (APPC) in Bihar, which is the average cost the utility pays to a generator to acquire one kWh of electricity. The APPC does not include costs of grid infrastructure or operation. The dotted lines show average utility revenue, with and without subsidies from the government, after cumulatively accounting for various sources of electricity loss. The official rate is for domestic consumers; industrial and other consumers may receive fewer subsidies on the margin.

subject to a universal access norm, but for which there exist a richer array of private alternatives.

When the power is switched off and villages go dark, the shortfall is often given technical terms like "load shedding," but at root it reflects a decision by the utility to sell less. The utility cannot—without frequent government bailouts—withstand the losses that would accrue from providing the 24/7/365 electricity common in high-income countries. The nonpayment social norm implies that consumers cannot get all the electricity for which they are willing to pay. The resulting poor supply harms residential and industrial consumers across the income distribution and acts as a brake on economic development. Treating electricity as a right therefore undercuts electricity access and reliability, holding back economic growth.

To illustrate this equilibrium, Figure 1 represents the pricing of electricity in Bihar, a large Indian state that has lately undertaken wide-scale electricity reforms. The government of Bihar has placed a high priority on widespread access to electricity and expresses this priority with a substantial subsidy for the production of electricity, equal to about 80 percent of the average cost of procuring a kilowatt-hour of electricity. Consumption subsidies are uncontroversial and indeed admirable policies for a government looking to increase consumption of a good that brings social and economic benefits to millions of people. However, a widespread social

belief that electricity is a right, combined with social and political constraints, makes it difficult to charge customers for electricity and nearly impossible to disconnect consumers who do not pay.

The result is high levels of nonpayment and theft. Of the electricity produced in Bihar, about 10 percent is lost during technical reasons during transmission. Another 20 percent is taken by illegal connections and not billed for at all. Of the remaining output that is billed for, about 15 percent results in unpaid bills. Furthermore, utilities are required to set tariffs significantly below the marginal cost of supply—in 2018, average household tariffs were about 4 cents per kilowatt-hour against a cost of supply of almost 7 cents per kilowatt-hour.

The gap between the cost of supply and revenue should, in principle, be made up by government subsidies. Subsidies, however, apply only to the actual quantity of power that consumers are billed on, not the additional quantity lost to theft and nonpayment. Furthermore, the payment of subsidies can sometimes be substantially delayed, widening the gap that utilities see between revenues and costs. As a result, even with a subsidy set at about 80 percent of the procurement cost of electricity, the providers of electricity in Bihar cannot cover their operating costs.[2] The old yarn that "we lose money on every unit, but will make up for it on volume" describes an untenable situation for any business. Therefore, the electricity distribution companies have no choice but to ration supply.

Reforms to improve payment performance and increase revenue collection therefore must occur alongside the push towards universal access. This fundamentally involves changing the norm that electricity is a right.

In Figure 1, the dotted red line shows that the average revenue obtained per unit of power supplied in Bihar, excluding subsidies and accounting for losses from billing inefficiencies, theft, and technical losses, is 2.3 cents per kilowatt-hour. This may be compared to the average cost of 6.8 cents per kilowatt-hour at which the Bihar utility bought power from generators in 2018–2019 as shown by the solid horizontal line (Prateek 2018). The shortfall is 4.5 cents per kilowatt-hour. The dotted blue line shows the portion of this shortfall that would appear on a utility's books, making the optimistic assumption that all subsidy reimbursements are made on-time and in-full. Although the difference here of 1.0 cents per kilowatt-hour is much smaller, this still means the utility would fall short of covering only the variable costs of electricity purchases—a far cry from attaining the cost-plus model based on which tariffs are set.

Over the remainder of this paper, we will refer to "losses" with the understanding that this is shorthand for the sum of subsidies and other losses including from theft.

[2]Note that this calculation, along with most of our analysis, considers variable costs. There are also subsidies for fixed costs, such as the costs of connecting a new consumer. Starting in September 2017, these fixed costs were waived entirely in Bihar for consumers below the poverty line, as part of the Indian government's Saubhagya scheme. Consumers above the poverty line were charged only a nominal fee. As with variable subsidies, fixed subsidies are supposed to be transferred to the utility, but may not arrive in a timely fashion.

In doing so, we will abstract away from the accounting distinction[3] between state-owned utilities and state governments.[4] We do not mean this shorthand to imply that governments and utilities have the same goals.

In the conclusion to this paper, we discuss various ways in which the descent into insolvency depicted in Figure 1 can be avoided. Only in this way can countries in the developing world get to the goal of universal 24/7/365 electricity.

Figure 2 illustrates what can happen to electricity markets when electricity is viewed as a right. Every point represents an electricity "feeder," which is a disaggregated level of the grid in Bihar that serves about 2,500 households and businesses on average. [5] The horizontal axis reports the "revenue rate," which we calculate as the ratio of the total revenue collected by the distribution companies at the feeder level to the revenue that would be collected if all power were paid for at prevailing tariffs.[6] In other words, the revenue rate measures distribution company efficiency and deviates from 1.0 because of technical losses, unbilled power, and unpaid bills—the steps down in Figure 1. The left vertical axis shows the daily hours of supply for each feeder, averaged by month; there are at most twelve data points for each feeder because the sample is one year long, but because of missing data, we observe months of data for each feeder on average.

[3]Figure 1 illustrates both the descent into insolvency that we have discussed and an accounting consideration that can be confusing when analyzing these markets. In the equilibrium we have described, a utility has two sources of revenue. The first of these are reimbursements from the government to cover losses due to tariff subsidies or waived connection costs. The second is the revenue obtained from consumers. Because distribution companies are typically owned by the government in developing countries, subsidy reimbursements merely transfer revenue shortfalls from the books of utilities to the government, without changing the fact that these costs are ultimately paid by taxpayers.

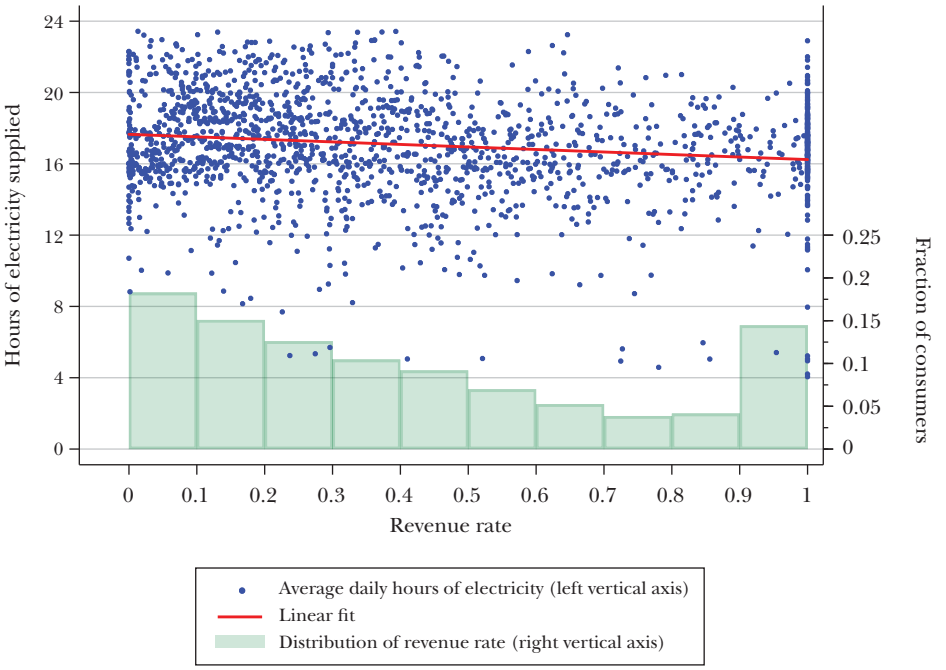
When subsidy reimbursements are not paid in full, the burden of these shortfalls falls on the utility. As an example, the World Bank tracked electricity subsidies in Bihar over a ten-year period from 2003–2013, finding that while subsidies booked rose by 17 percent year-on-year, reimbursements grew by only 12 percent (Pargal and Banerjee 2014). Over a ten-year period, the authors estimate a shortfall of a staggering $7.5 billion (in 2013 US dollars).

[4]In addition to this abstraction, we make certain simplifications in Figure 1 to aid exposition. First, in practice, utility losses are a weighted average of consumer-type specific losses. Thus, in a complete accounting, a similar pair of lines should be separately drawn for every consumer class. Here, we use numbers for the first block of domestic (DS-1 and DS-2) consumer tariffs, applicable to the households who consume between 0 and 50 kWh per month. The second simplification comes from our discussion of subsidy transfers. This is because in practice transfers are based not on the cost of supply, as we describe in the main text, but on consumer-type specific differences between regulatory tariffs and subsidized tariffs. This creates a network of cross-subsidies that are set such that in aggregate utilities are made whole (discounting technical losses, theft, and nonpayment). Describing this full accounting structure is beyond the scope of this paper.

[5]The figure is based on 2017–2018 data from a sample of 172 feeders that are representative of the population of feeders in eight districts of Bihar, excluding district headquarters. This paper is largely focused around the failure of developing countries to provide electricity *outside* large cities. In large urban areas, the electricity access and reliability problems are much closer to being solved. These detailed administrative data form part of a separate ongoing randomized experiment are being conducted by the authors and documented in the social science registry (AEARCTR-0000479).

[6]These tariffs may be different for different types of consumers, so the denominator of the revenue rate is not a constant multiplied by energy but rather depends on the consumer mix served.

Figure 2
Hours of Electricity versus Fraction of Revenue Collected for Selected Feeders in Bihar



Source: Bihar Electrification Project.
Note: This figure shows the hours of electricity supplied to different areas each day (left vertical axis) against the share of the cost of electricity that each area pays (horizontal axis) for 172 feeders in north and south Bihar for the period between May 2017–April 2018. Feeders are a representative subset of the population of feeders in eight districts of Bihar, excluding district headquarters, and thus serve primarily rural areas and small towns. Each observation reflects revenue and supply in a particular month, for months in which both variables are observed in the distribution companies' administrative data at the feeder level. The revenue rate is calculated as the total payments for electricity divided by the value (at publishing post-subsidy tariff rates) of energy injected at the feeder. The revenue rate therefore ranges between zero and one for areas that pay none or all of their bills, respectively.

Three facts about the retail electricity market in rural or small-town Bihar emerge from this figure. First, the supply of electricity is heavily rationed and variable. In this sample, *no consumer* gets 24 hours of electricity every day; on average, consumers receive about 17 hours a day, and some areas get only 12 hours a day. Further, there is great temporal variation in supply within a given feeder that likely imposes costs on customers, beyond the costs associated with having less than 24 hours of service on average.

Second, revenue rates for electricity are low and variable. Our revenue rate, which measures the all-in ratio of actual revenue from customers to the value of energy injected (at published rates), is smeared out along the horizontal axis. Some areas (to the right of the figure) are paying the full share of energy value, but many

more customers pay a share of less than 0.20, on the left. The average revenue rate in this sample is only 38 percent and 75 percent of feeders pay less than two-thirds of the value of energy injection. These losses are much higher than for Bihar as a whole, in large part because this sample of feeders covers largely rural areas as well as some small urban or peri-urban portions of these districts, and thus is not representative of Bihar.[7] It excludes large towns where the utility collects over 90 percent of what it is owed. However, this does underscore the challenges in expanding access to poor and largely rural populations, which is the primary focus of this paper.

Third, but most striking, the scatter plot shows that the relationship between how much supply people receive and how much they pay is slightly *negative*. Areas that pay for the entire cost of power tend to get a little less power than areas that pay nothing. This is evident in the solid red line of best fit that is slightly negative in slope, contrasting sharply with the market for a typical good where consumers who pay more would tend to get more.

To put these patterns in further context, consider how this graph would appear for power consumption in a high-income country. Outage rates are extremely low, so all areas would be at or extremely close to 24 hours of supply. Loss and nonpayment rates are also low, so all areas would have a revenue rate of almost one. Consequently, these scattered points would collapse to a single point in the northeast corner. In contrast, the pattern we observe in Bihar illustrates a situation where the link between payment and supply has been severed.

At the heart of this study is the recognition that the problems plaguing Bihar's electricity markets are shared, to a greater or lesser degree, by many developing countries. As we will discuss, the consequences of providing electricity regardless of payment—large subsidies, high rates of theft and nonpayment, indebted distribution companies, restricted access, and frequent blackouts for paying customers—are visible in many countries. We argue that a key part of the problem, perhaps the key part, is the same: unlike in other domains, when public provision of electricity collapses, households lack reasonably equivalent private substitutes. Electricity is a natural monopoly: average cost is decreasing for all quantities, so it is efficient to have one grid. Households do substitute, but they substitute to the equivalent of electricity autarky—off-grid diesel generators or solar panels that cost far more than grid electricity and provide smaller loads (Burgess et al. 2019). In the equilibrium we are describing, there is therefore a pent-up demand for electricity from consumers who are *able and willing to pay for it*, meaning that socially beneficial transactions simply do not take place. The rationing away of electricity from these consumers, on the intensive margin of hours of supply per day, is also mirrored by the rationing in access to electricity on the extensive margin.

[7] The statewide average revenue rate under our methodology can be read off the red line in Figure 1, that is, 2.3 divided by 3.8 or about 60 percent. As a detail, the calculation we present uses only data from residential consumers. The overall losses of the utility require a similar accounting across other consumer types and was reported in 2018–2019 at 36 percent, implying a revenue rate of 64 percent that is roughly consistent with our calculation.

The consequences of this state of affairs for development are likely severe. Electricity is an essential input for production, consumption, communication, and finance. Indeed, there are no examples of societies that have reached high living standards without consuming high levels of electricity.[8] Confronting the global energy access and reliability problem will therefore be a key means of encouraging future growth and poverty reduction.

We lay out this argument in several stages. The following section contrasts electricity losses in low- versus high-income countries, thus providing a picture of the global extent of the problem. To understand why electricity utilities in developing countries ration their product, we next write down a graphical model. We then do a deep dive into microdata from the Indian state of Bihar to unpack in detail the mechanisms underlying the dynamics of the electricity market there. Our study site thus serves as a detailed illustration of issues we argue are widespread in developing countries. As scaffolding for this analysis, we use the four-step structure described above: 1) thanks to social norms, consumers view electricity as a right, and subsidies, theft, and nonpayment are tolerated; 2) electricity distribution becomes loss-making; 3) distribution companies ration electricity supply; and 4) supply and payment are delinked. In the conclusion, we offer some suggestions for reforms. These recommendations apply not just to Bihar but also to countries across the world seeking to obtain universal 24/7/365 electricity and the economic growth that it facilitates.

Electricity Losses around the World

A key insight from this paper is that two energy worlds coexist, one where consumers enjoy universal access to electricity 24 hours a day and another where many consumers are not on the grid and those who are connected suffer irregular supply. Panel A of Table 1 shows the differences in these worlds through statistics on electricity use for countries classified by income into four broad categories.

In some respects, the two energy worlds differ only in degree, in a way that may be taken purely as intrinsic to the differences in income levels between poor and rich countries. Electricity consumption in low-income countries is a negligible 1 percent of that in the United States; thus, world inequality in electricity is larger than income inequality. All consumers in high-income countries have electricity, whereas only 35 percent do in the low-income countries. It is plausible that some of these unconnected poor have low demand for electricity and it would lower social surplus to connect them to the grid (Lee, Miguel, and Wolfram 2019). However, we

[8] There is substantial evidence that access to reliable electricity can increase business profits, firm entry, labor productivity, and other inputs to growth (Allcott, Collard-Wexler, and O'Connell 2016; Dinkelman 2011; Kassem 2018; Fried and Lagakos 2017; Fried and Lagakos 2019; Moneke 2019). Electricity appears not only to boost output and labor supply in the short run but to raise long-run levels of productivity (Lipscomb, Mobarak, and Barham 2013). See also the companion paper in this symposium by Lee, Miguel, and Wolfram.

Table 1
Key Electricity Summary Statistics by Income Level

Quartile	*Lowest*	*Lower middle*	*Upper middle*	*Highest*
A: World Electricity Overview				
Population (millions)	619	2,972	2,568	1,165
GDP per capita in 2016 (% of US)	2.9	10.7	26.7	79.8
Electricity consumption per capita (% of US)	1.1	5.9	27.2	69.9
Connection to grid (%)	34.9	83.6	99.4	100.0
T&D loss (%)	22.8	16.2	9.6	6.1
Firm losses due to outages (% of output)	8.7	6.6	2.1	1.6
B: Pricing in Selected Countries				
Mean monthly residential consumption per electrified household (kWh)	98	103	162	574
Mean price at mean consumption level (US cents/kWh)	3.6	6.3	7.6	18.8
Mean power purchase cost (US cents/kWh)	6.4	7.2	6.6	6.2
Power purchase cost after T&D loss adjustment (US cents/kWh)	7.8	8.3	7.5	6.6
Mean price less adj. power purchase cost (US cents/kWh)	–4.2	–2.0	0.1	12.2

Source: World Bank, IEA,World Energy Council, country sources.
Note: This table shows electricity variables for four income categories of countries, using the 2018 World Bank thresholds of 2016 GNI per capita of ($1,005; $3,955; $12,235). Panel A displays population-weighted averages for all countries in each income category. In Panel B, the sample consists of the ten largest countries worldwide by population as well as the three most populous in each WB income category: Ethiopia, DR Congo, and Tanzania (lowest); Bangladesh, India, Indonesia, Nigeria, Pakistan, and the Philippines (lower middle); Brazil, China, Mexico, and Russia (upper middle); and France, Japan, and the United States (highest). In Panel B, the first row is an unweighted average across selected countries. In other rows, average prices and costs are weighted by utility customers for the three largest utilities within selected countries and unweighted across selected countries. The individual country sources include government statistics websites and specific utilities' websites.

will argue that another reason for low access is due to electricity being treated as a right on the supply side of the market and that this reduces welfare by preventing socially beneficial transactions from taking place.

Certain differences between the energy worlds do not seem intrinsic to income. For example, transmission and distribution (T&D) losses are about four times higher in the low-income countries as in the high-income countries (22.8 versus 6.1 percent). Yet the technologies used for distribution are largely the same everywhere: although the levels of investment or structure of the distribution network may be different, there is no way to justify a fourfold difference in losses on technical grounds alone. The divergence must be generated at least in part by social or institutional factors that vary across countries, such as—we argue in this paper—by social norms around electricity provision that contribute to poor bill payment rates and higher losses in low-income countries.

Low-income countries also price power below cost. Panel B of Table 1 shows that in low-income countries, utilities pay a mean power purchase cost of 6.4 cents

per kWh and charge customers 3.6 cents per kWh for the same power. If we inflate power purchase costs by transmission and distribution losses, since utilities have to buy more input power to make up for the power they lose, then the input cost is 7.8 cents per kWh. Thus, the average utility in a poor country makes 46 cents per dollar of input cost—and even this calculation is optimistic, as it excludes the nonenergy variable costs of distribution and commercial losses from power billed but not paid for. Utilities in the lower-middle countries ranked by income also price power below cost (second column). But in the high-income counties, the average price paid by consumers for electricity is roughly three times higher than the mean power purchase cost for utilities (18.8 cents relative to 6.6 cents per kilowatt-hour), presumably reflecting the need to cover fixed costs and distribution company profits.

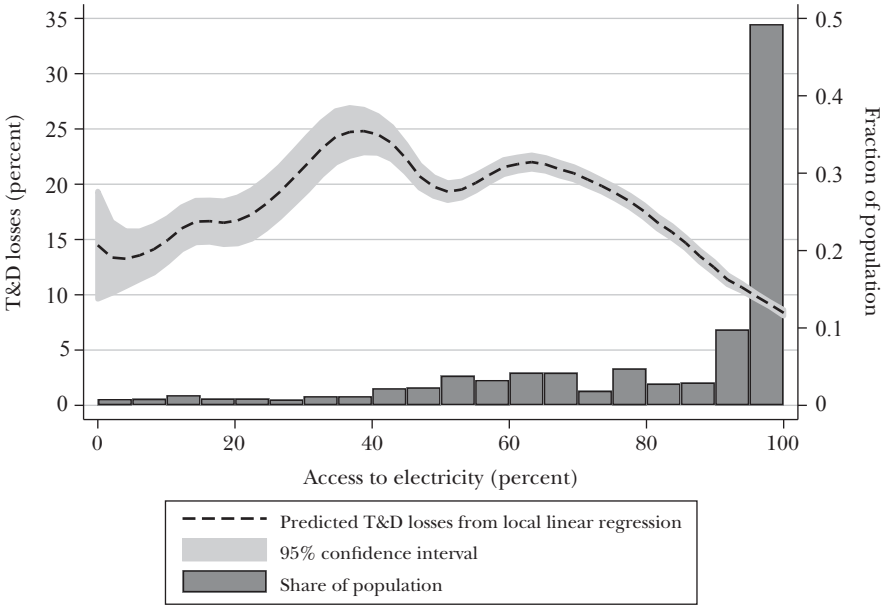
The last row of Table 1, which gives the difference between the average price the utility is paid and the average amount it must pay to generators, is therefore an upper bound on utility profit per kilowatt-hour. The difference is negative for low- and lower-middle-income countries, suggesting that *utilities in poorer countries do not cover even the raw costs of power acquired from generators.* Utilities in these income brackets are therefore unprofitable and must be supported by government subsidies and grants. Including commercial losses, like power that is billed but not paid for, would further inflate these losses. If we were to include the high fixed costs of grid infrastructure, lower-income countries would have no hope of making electricity provision profitable; even in the richest countries, utilities—as regulated monopolies—tend to barely break even after fixed network costs. [9]

We argue that these differences are not purely due to poverty, because it is conceptually possible that low-income countries would manage well-run grids, where electricity was not lost and people just used a low quantity due to low demand for electricity at low incomes. But that is not the case—people do use little electricity, but losses are high. The average price shown in Panel B is also lower in these countries, in large part due to subsidies, further increasing the gap between costs and revenues. In poor countries, therefore, utilities lose money on every unit of electricity they sell.

An implication of losing money when supplying electricity is that attempts to expand access to electricity in low-income countries will increase losses. Transmission and distribution (T&D) losses refer to the share of power generated that goes unbilled (as opposed to commercial losses, which are power billed, but not paid for). As mentioned earlier, a small amount of power (around 5 to 10 percent) is lost for unavoidable technical reasons known as "line losses." Losses much above this level come from unregistered and registered consumers hooking onto distribution wires, unmetered power, meter tampering, or other forms of theft. In Figure 3, we plot T&D losses against percent access to electricity (the share of the population

[9] Of course, a lack of profitability in the electricity market does not itself imply that these subsidies cannot be welfare-enhancing. Rather, our point is that tackling the problems associated with electricity being viewed as a right could increase welfare by enabling socially beneficial transactions to take place.

Figure 3
Access to Electricity and Transmission and Distribution (T&D) Losses



Source: World Bank.
Note: Each point represents one country and year for all 142 countries and years from 1990–2014 for which data are available. The local linear regression and histogram of access to electricity are both weighted by country population. T&D losses are defined as the percent of electricity generated by all power sources (in kWh) that is not billed to any consumer. Access data were originally gathered from household surveys, and T&D data are originally from national energy agencies.

with a grid connection) from 1990–2014 for all countries with available data and fit a nonparametric regression. Data from 142 countries are included with 125 of these countries having nonmissing data in all 25 years.

Figure 3 plots the result, an inverse-U shape where losses rise and then fall in access. For countries in years where access to electricity is very low, transmission and distribution (T&D) losses are high, but losses actually increase further as access expands before falling again as access approaches 100 percent. In other words, countries which are trying to expand distribution (for example, into the countryside) face the highest rates of nonpayment for electricity. At the peak of the curve, countries with about 40 percent access to electricity on average lose 25 percent of their power before it is billed to any consumer. Many states in Nigeria and India, among other places, exhibit T&D losses of 33 percent or more (Government of India Ministry of Power 2019; Nigerian Electricity Regulatory Commission 2019), which implies that the electrical utility is giving away one in three units of electricity for free. Losses then decline as payment norms are established and enforced for richer countries at higher levels of access.

We call the inverse-U relationship in Figure 3 an electricity Kuznets curve. The original Kuznets (1955) curve documented an increase and subsequent decrease in inequality as a function of income. In this version, electricity distribution companies see losses initially increase as they move up from low levels of access, but then decline as access becomes more widespread. It should be noted that our electricity Kuznets curve, like the original, relies on data from several countries and thus does not perfectly describe the path of a single country that attempts to increase access. Nevertheless, it makes clear that because electricity distribution is loss-making, governments making efforts to reach more or all of the population will for some period face higher losses.

Mechanisms: A Model to Explain Electricity Rationing

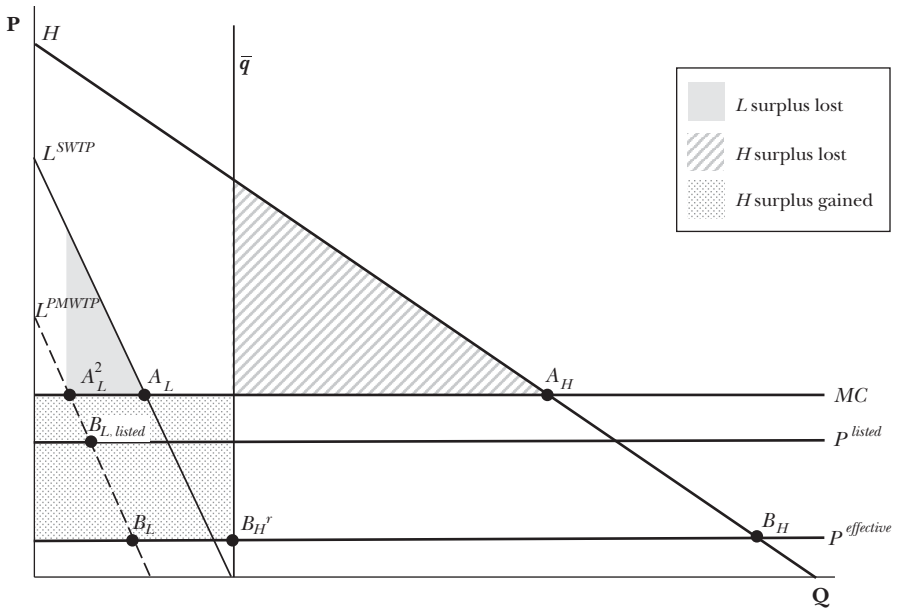
In this section, we examine the mechanisms via which viewing electricity as a right combined with its nonexcludability causes utilities to ration supply. Consider the case of two types of consumers, H (high income) and L (low income), illustrated in Figure 4. The demand curve of the H type is labeled H. The demand curve for the L type is denoted by L^{PMWTP}, which reflects their private marginal willingness to pay (which of course depends on their ability to pay, given low incomes).

The treatment of electricity as a right means that society values each unit of consumption by the poor above their own willingness-to-pay. Such a belief could arise for a variety of reasons, including because the state finds it dignified that the poor have light in their homes, due to market failures like credit constraints that limit the poor's ability to pay their full private valuation, or because there are network externalities. This belief is reflected by L^{SWTP}, which represents societal willingness-to-pay of L consumers, lying above L^{PMWTP}. Indeed, the idea that social willingness-to-pay is above private willingness-to-pay—as highlighted by statements from Indian politicians and international aid organizations in campaigns for universal electrification—is a motivation for why public provision of electricity exists in the first place (Banerjee 1997).

At marginal cost MC, the efficient quantities of consumption are A_L and A_H. However, if the state set power prices at this level, L consumers would only consume at $A_L^?$, according to their own private willingness-to-pay, generating a deadweight loss relative to the social optimum determined by L^{SWTP}. This deadweight loss is denoted by the solid grey triangle in the figure and labeled as L surplus lost. Marginal cost pricing fails to deliver the social optimum here because society places a value on L consumers' consumption that is over and above their own valuation.

One option here is for the state to set a lower price, P^{listed}, below marginal cost to encourage additional consumption. At this price, the L types would increase consumption to $B_{L,listed}$. The state would lose $(MC - P^{listed})\ B_{L,listed}$ in subsidies, and the poor consume closer to the socially efficient quantity. But notice that this subsidy

Figure 4
A Mechanism for Electricity Rationing



Note: This model illustrates how a perceived right to electricity, combined with the nonexcludability of electricity, leads utilities to ration supply. High-income consumers have demand H; since society places an additional value on electricity consumption by the poor, social willingness-to-pay L^{SWTP} for low-income consumers exceeds their private demand L^{PMWTP}. Pricing electricity at marginal cost MC leads to deadweight loss for L consumers, since they consume at A_L^2 instead of the efficient quantity A_L. Pricing at a slightly lower P^{listed}, through a subsidy to L-types, increases consumption to $B_{L,\ listed}$, which is still below the efficient quantity. However, when social norms and nonexcludability result in nonpayment, the effective price falls to $P^{effective}$ and L-types can consume at B_L, close to efficient A_L. Faced with average losses of $(MC\text{-}P^{effective})$, the utility curtails supply to $\bar{q}$. In this equilibrium, H-types consume at B_h^r, and may face losses in surplus.

does not fully move electricity consumption of L types to the socially preferred level A_L.

However, a combination of the social norm that electricity is a right and the costs of making electricity excludable limits the ability of the state to collect revenue. The *effective* price that consumers face is therefore much lower, at $P^{Effective}$, and the poor consume B_L at this lower price. The state makes a larger loss of $(MC - P^{Effective})\ B_L$, but the poor consume even closer to the efficient quantity—that is, B_L is closer to A_L than $B_{L,\ listed}$.

Moreover, if this very low price were applied to both types, the high-income H consumers would use "too much" and consume at point B_H. The loss associated with serving these consumers is larger than the loss from serving L types because the H types are richer and consume so much more. Furthermore, the state does not value the excess of their consumption over the efficient level and would make enormous losses of $(MC - P^{Effective})\ B_H$ on their supply.

One solution to this problem might be to use block-rate tariffs to charge a lower price for the first increments of electricity consumers and a higher price for additional quantities. In this case, the higher income H and lower income L consumers would not face the same marginal price. The trouble is that the high costs of making electricity excludable, combined with widespread nonpayment inevitably arising from the social norm, mean that in practice, the state is not able to price discriminate between H and L types. Therefore, the effective price is indeed low for everyone (we provide some empirical evidence for this assumption later in the discussion).

However, an electricity provider under severe budgetary pressure has another instrument at its disposal: quantity rationing. One option is to limit supply to H types to the efficient level of A_H. However, at this level, the state will still make large losses and may not value the surplus of the H types at all. Furthermore, if there were many consumer types, the effective price may be very low, and the state is limited by its budget constraint. There is no reason to think utilities will be solvent with only the small degree of rationing to A_H.

Thus, in order to keep enough funds to continue supplying all the types together, the electricity provider may ration further to a point like $\bar{q}$. At $\bar{q}$, L type consumers use a quantity close to their efficient quantity and would not want to pay much higher prices for the small gain in gross surplus that pricing at cost would bring them. But H consumers have been cut back sharply to B'_H. These consumers are using much less than the efficient level of power; the well-off farmer will not have a refrigerator, for example, or a rural metal shop will continue to use only hand tools.

Despite the fact that the high-income H types are paying low prices, their loss of surplus may be great enough that they would prefer a regime with full supply and prices raised to cover costs. The H consumer has gained the dotted area in the figure labeled "H surplus gained," since power is so cheap. However, the H consumer has lost the shaded triangle, "H surplus lost," which would have been part of consumer surplus with marginal cost pricing and no rationing. The lost surplus from rationing may well outweigh the gain from high prices; the sign of this trade-off is ambiguous. What is clear is that, due to rationing, the marginal unit of electricity for these high-income H consumers is valued far above the unit cost that they pay. Yet despite this, H consumers cannot buy more electricity.

The Consequences of Treating Electricity as a Right

This section uses empirical data to walk through the different steps that begin with treating electricity as a right, and end with crippling electricity rationing. The facts that we will document are (1) energy is viewed as a right; (2) this results in subsidies, theft, and distribution companies losing money; (3) which leads to the rationing of supply; and (4) the delinking of supply from payment. All these four factors erode payment incentives for private consumers, reinforcing the viewpoint we started with, namely that electricity is a right and not a private good.

Table 2

Customer Beliefs about Enforcement in Rural and Small-Town Bihar, India

(*Percentage responses to: If you did X, how likely would it be that you would incur any penalty from the distribution company?*)

	Likely	*Neutral*	*Unlikely*
Paying your bill late	10.1	13.6	76.3
Modifying your meter	7.9	18.2	73.9
Having an informal hooked connection	7.6	14.4	78.0
Bribing electricity officials	12.2	24.5	63.3

Source: Bihar Electrification Project endline household survey, May–August 2017.
Note: Responses are from a survey of 7,071 households in rural and small-town Bihar. Modifying a meter, having an informal hooked connection, and bribing officials all prevent a utility from observing actual electricity consumed and therefore constitute power theft.

Our evidence in this section consists mainly of microdata from Bihar, including monthly bills for over 5 million households and businesses, as well as accompanying survey data. We also incorporate some international evidence. Electricity utilities in many developing countries share remarkably similar institutional setups to those observed in Bihar. Moreover, as we have documented earlier, high levels of subsidies, theft, and nonpayment leading to high electricity losses and rationing characterize the situation in a range of developing countries. As a result, the cycle outlined below may help us to understand why restricted access and unreliable supply characterize many electricity markets.[10]

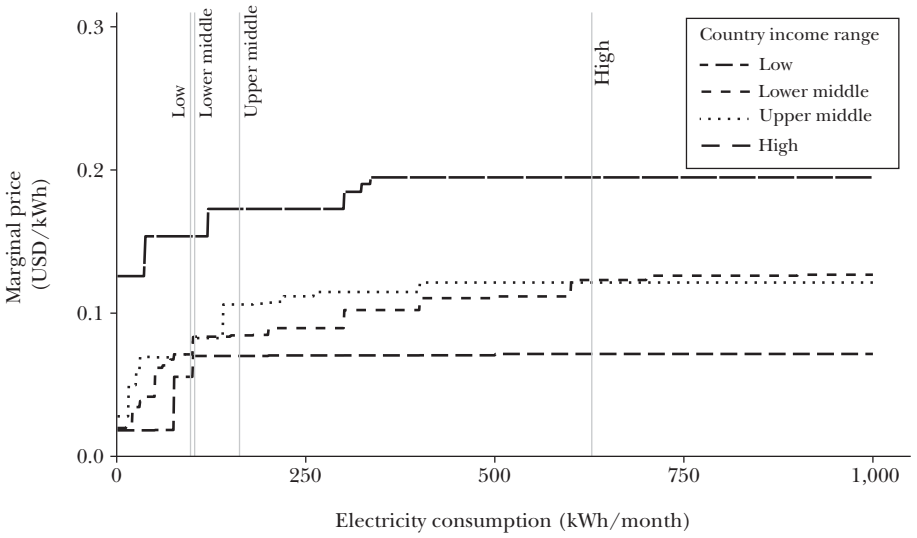
Step 1: Electricity Is Seen as a Right

Table 2 documents that the vast majority of customers in Bihar expect *no penalty* from paying a bill late, illegally hooking into the grid, wiring around a meter, or even bribing electricity officials to avoid payment. These attitudes are in stark contrast to how the same consumers view payment for private goods like cellphones. It is debatable whether cellphones are more important than electricity, but in Bihar we find that the poor spend three times *more* on cellphones than they do on electricity (1.7 versus 0.6 percent of total expenditure). These small expenditure shares for electricity suggest the lack of payment for electricity stems not from an inability to pay, but rather the norm of nonpayment.

A second piece of evidence comes from how poor countries set electricity prices. Figure 5 plots the published marginal price of each kilowatt-hour (kWh) of electricity, averaged across countries within a World Bank income group. The vertical lines in the figure indicate the average level of consumption for consumers

[10] Clearly other factors beyond those that we examine here may contribute to this state of affairs. For example, countries such as Nigeria do not have enough power plants (installed capacity), and in Latin America power sector unions have been known to curtail electricity provision to extract rents.

Figure 5
Explicit Subsidies in the Marginal Price of Power



Source: Electricity tariff (rate) schedules published by selected utilities.
Note: The graph shows the published marginal price of an additional kilowatt-hour (kWh) of power for selected countries within a 2018 World Bank income group. In general, the cheapest available domestic/household rate is used. Selected countries are in the union of the three largest countries by population in each income group and the ten largest countries worldwide. We construct each country's price schedule separately and compute unweighted average prices at each kWh level. For countries with multiple rate schedules, we use the three largest utilities by number of customers (five for India) and take a weighted average by customer count to construct the country schedule. Utilities sometimes adjust fixed charges or the marginal price on previous units when a consumption threshold is exceeded; those one-time increases in the marginal price are not included.

in each group of countries. Utilities everywhere charge less for consumers who use small amounts of power. The price of power on the first step is low and then steps up for greater consumption. Across our sample of 30 utilities in 16 countries, almost every utility charges less for the first few kilowatt-hours than for remaining units. The first steps of such tariffs are sometimes explicitly called "lifeline" tariffs, suggesting that social norms around access for the poorest can affect the utility's decision to give away electricity below cost.

While the marginal price of purchasing electricity increases with consumption in both low- and high-income countries, the difference is much greater in low-income countries (a factor of 3.9 rather than a factor of 1.5 in high-income countries). Moreover, because poor consumers use less, many more people are consuming power at the highly subsidized initial rates. Even at higher energy consumption levels, electricity rates in low-income countries tend to be much lower than in rich countries. It may be that fixed costs of distribution are also lower in poor countries, but this does not seem to be the main story, as the highest tariff steps are still below the cost of power purchase alone (as

shown earlier in Table 1). The pricing of power below cost means that electricity distribution companies are set to lose money even if every consumer paid their bills.[11]

Beyond subsidies and technical losses, the next two steps into insolvency come from power that is not billed and nonpayment of bills (illustrated earlier in Figure 1). Unbilled power is often referred to as "theft" because a significant fraction of unbilled power may be stolen through measures such as hooking wires illegally to overhead lines. However, some power may not be billed because of billing inefficiencies on the part of the utility. We have shown that transmission and distribution losses in the electricity system are higher in poor countries, but there is not comparable data breaking down unbilled power and nonpayment across a range of countries. From our data on Bihar, however, we can look more carefully at how power is lost and who does not pay for it.

We showed earlier that the revenue rate—that is, the ratio of payment that is collected for electricity to the collections that would occur if all consumers were properly charged—is surprisingly low in rural and small-town Bihar (illustrated earlier in Figure 2). Low collection could be due to outright theft, which would show up as power that is not billed. Here, we show that a surprisingly large part of losses stem from known, formal customers not paying their bills.

Figure 6 utilizes administrative billing data from households in rural and small-town Bihar and plots the bill payment rate against monthly electricity consumed, averaged across each month in 2018 for the subset of households that receive bills. The payment rate conditional on receiving a bill (dashed line) is roughly flat across the consumption distribution, or even slightly declining, implying that bigger consumers are just as delinquent on their electricity bills as smaller ones. More than half of collection losses are due to nonpayment by consumers using over 100 kilowatt-hours per month (one minus the dotted line), though the histogram shows that they are a small subset of domestic consumers in our sample.[12]

The finding that nonpayment conditional on being a formal customer and receiving a bill is both high and constant across the distribution of consumption suggests that *de facto* low effective prices are an accepted and agreed-upon policy of the state. These customers are administratively known to the utility—clearly identifiable in their data—but are not paying and remain connected customers, while piling up debt month after month.

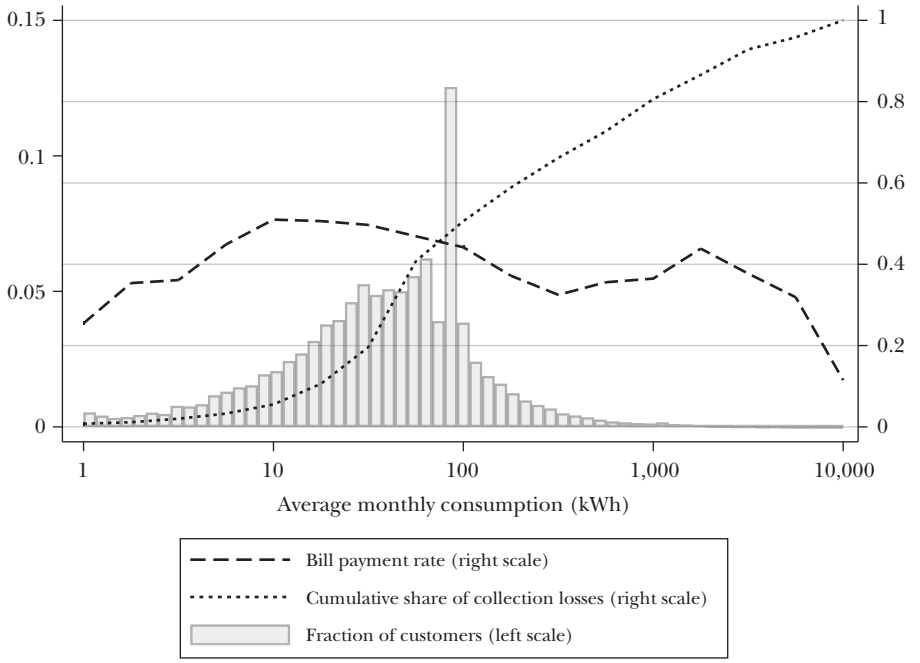
Step 2: Electricity Distribution Is Loss-Making

Thanks to subsidies, theft, and nonpayment, governments in poor countries lose a lot of money. The utility would be able to recover 85 percent of procurement

[11] Recall that our use of the term "distribution companies" is shorthand for the combination of the government and the state-owned company—if subsidies are reimbursed in full to the distribution companies, these losses simply move to the books of the state government.

[12] Data consist of individual customer bills from October 2017 to June 2018 from feeders in five districts of Bihar (out of the eight districts covered in Figure 2), all excluding district headquarters. Since payment rates in large urban centers are very high, this figure is indicative of small-town and rural consumers.

Figure 6
Bill Payment Rates for Selected Feeders in Bihar, India



Source: Bihar Electrification Project.
Note: The graph shows average bill payment rates by kilowatt-hour (kWh) consumption level, as well as the share of collection (bill payment) losses accounted for by consumers below that level. Only consumers who are actually billed are included. The bill payment rate equals revenue received as a share of the billed amount and therefore does not account for unbilled power (theft). Consumption brackets are 0–0.25 log10 kWh, 0.25–0.50 log10 kWh, etc. Customers with monthly household consumption above 100 kWh account for half of all collection losses. Data consist of electricity bills from October 2017 to June 2018 from 1.49 million unique customers. All customers are from one of five districts of North Bihar, out of the eight covered in Figure 2, all excluding district headquarters. Since payment rates in large urban centers are quite high, this figure is indicative of consumption and payment behavior of small-town and rural consumers.

costs net of subsidy transfers, but Bihar as a whole recovers about 34 percent of costs. In other words, across a population of 100 million people, payments from consumers cover less than half of the cost of power. As we discussed earlier, these losses are distributed in different ways between governments and the distribution companies they regulate, but the key point is that because revenues are less than costs on a per-unit basis, expansions of output ultimately require increases in tax revenues.

The problem with a power sector reliant on debt and subsidies is that at some point, electrical utilities run out of money. A number of countries have run up substantial power sector debt: in some cases, enough to have macroeconomic implications. In Pakistan, accumulated electricity debt is almost 4 percent of GDP (Babar 2018). India was facing stressed power debts of $62.5 billion in mid-2018,

amounting to 2.4 percent of GDP (Engelmeier 2015). These debts, including $30 billion of loans owed directly to distribution companies, threatened to instigate a financial crisis. Underscoring the speed at which power debt can accumulate, it should be noted that India's current distribution company debts exist in spite of a $42 billion central government bailout in 2016 and 2017 to save states from insolvency, which followed earlier bailouts in 2011 and 2002 (PTI 2018). There appears to be a 7–10-year cycle of power sector bailouts in India.

Power sector debt in Nigeria has also been reported to scare off private investments in generation and in Ghana leads to power rationing (Akwagyriram and Carsten 2018; GhanaWeb 2018). During the Puerto Rico debt crisis in the United States, the state-run power utility owed $9 billion in debt, in part because it gave free power for years to government-owned agencies and businesses (Walsh 2016). The implications of a loss-making electricity sector for the wider spending objectives of government are therefore nontrivial.

Step 3: Distribution Companies Ration Supply

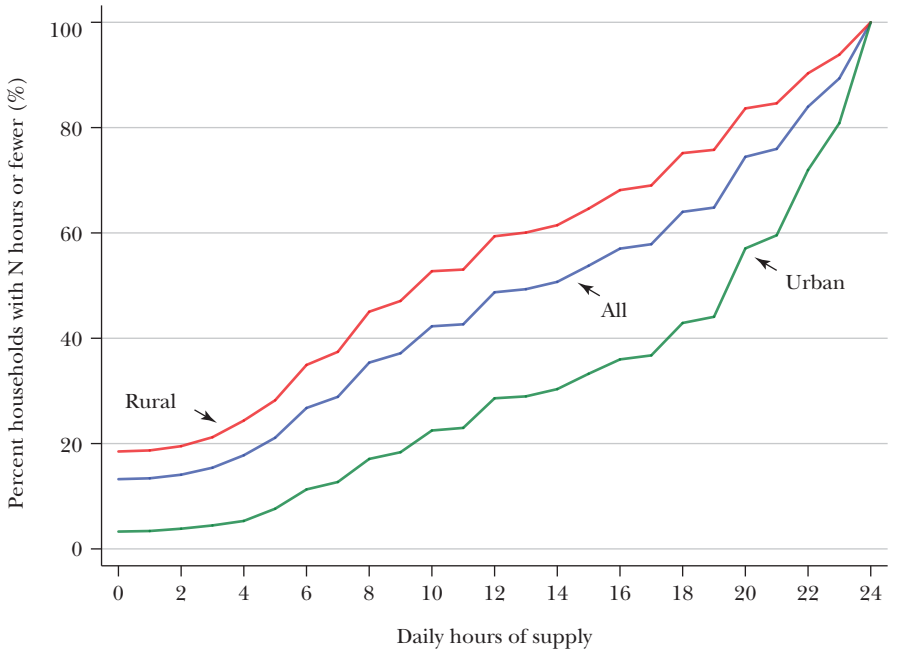
When electricity utilities in low-income countries are losing money on each unit sold and unable to shut down due to their public mandate, the only remaining option is to sell less by purposefully restricting supply. In practice, quantity is rationed by restricting the hours of supply on the grid. This practice is given fancy names, like "load shedding," but at its core it is a company choosing to sell less of its product even though some of its customers are willing to pay more than the cost of supply.

India is the largest country by population that faces electricity rationing. Figure 7 gives the distribution of daily hours of supply across the country in 2012. In rural areas, the median household received under 10 hours of electricity per day. Urban areas received over 19 hours. These numbers have improved, but only a small proportion of the population enjoys 24-hour electricity.

Rationing in India is not due to any absolute scarcity of capacity for generating electricity. In 2012, the year the data for Figure 7 was collected, coal plant utilization in India was under 70 percent, and in 2018, it is 55 percent. From the point of view of Bihar, which uses a small share of India's power and is connected to a national grid, an essentially perfectly elastic supply of power is available at a reasonable cost on wholesale power markets. *There is no shortage of power; rather, it is the inability to fully recoup the costs of electricity use that prevents India from providing a 24/7/365 flow of electricity to all of its citizens.* There is therefore a misallocation of power in India so that many people cannot buy the power they want. The same is true of places like Pakistan, which is now backing down from Chinese-funded coal plants, and Ethiopia, which benefits from abundant hydropower resources.

Developing countries also experience physical shortages of power and blackouts due to exogenous technical shocks, like the overheating of a transmission line. These shortages are best thought of as long-run consequences of rationing. When a high number of such shortages occur, the ultimate cause is mispricing and losses

Figure 7
Cumulative Distribution Function of Hours of Power Supply in India



Source: IHDS 2011–2012.
Note: This figure shows the empirical cumulative distribution function of the hours of electricity supply reported by rural (red), urban (green), and all (blue) households in the India Human Development Survey, 2011–2012. Households reporting no electricity have been considered to receive zero hours of daily supply. At each point in the distribution, rural households have fewer hours of electricity than urban households. The median urban household receives over 19 hours of electricity per day, while the median rural household receives under 10 hours of electricity per day.

associated with electricity distribution. A lack of revenue flowing into the sector undercuts investment in generation and transmission. Ghana's most recent power crisis provides an example of this type (Kumi 2017). McRae (2015) shows how utilities serving a population of poor consumers may provide a low-quality supply if consumers are unwilling to pay for power, but the utility continues to serve these customers because its losses are covered by subsidies it receives from the state.

Step 4: Supply and Payment Become Delinked

The overwhelming impression from Figure 2 presented earlier is that how much a community pays bears little relationship to how many hours of electricity it receives. Supply and payment have become delinked in the market for electricity in Bihar. In principle, a utility could ration judiciously by area. At a higher level of aggregation—say, at the feeder level—electricity is perfectly excludable. For example, a utility could give 24 hours of electricity to areas with high payment rates

and less to those that do not. Alternatively, a utility could give more power to areas that value electricity more highly, perhaps because they include more businesses or public facilities like hospitals.

Even if a utility is physically able to cut off a group of delinquent customers, the right that citizens feel to electricity is a social and political concept, not a technical one. This insight helps to explain the pattern seen in Figure 2 where that rationing of electricity bears no relation to the payment rates of different areas. In continuing to keep the lights on for nonpayers, the utility reveals that it is constrained from acting like a profit-maximizing business.

A narrow interpretation is that utilities do not take this approach because of a technical limitation that they cannot ration finely enough. Recall Figure 2, where each point represents a feeder that serves a community, not individual people. Thus, even if supply were linked to feeder-level payment, customers end up being accountable for the power theft of their neighbors. A public goods problem arises here via electricity payments, where people are unwilling to pay if the result is to make it easier for their neighbors to receive electricity without paying. If the utility were able to selectively and inexpensively cut off individual consumers who do not pay, perhaps it would do so, and indeed this is common practice in the developed nations.

A broader interpretation is that under the social norm that energy is a right, the allocation of power is no longer being done on purely economic grounds, just as the pricing of power is not. If citizens engage in protests regarding a poor supply of electricity, or equivalently if government or company officials urge action to increase supply, a utility will need to take such pressures into account when making its supply decisions because ultimately it is the government that underwrites the cost of electricity provision. A growing literature documents influences of this nature on electricity supply (Mahadevan 2019; Asher and Novosad 2017; Baskaran, Min, and Uppal 2015; Shaukat 2018).

When power is supplied or rationed on criteria other than economic return and payment, consumers have little incentive to pay for electricity. They quickly learn that the way to get more power for their communities is to appeal to the local electricity grid operator, company officials, or elected representatives. The unpredictable supply makes many consumers feel that they are being treated unfairly and additionally weakens incentives to pay. Consequently, the four-part cycle we have described will repeat itself.

Conclusion and Possible Reforms

When a social norm develops that electricity is a right, firms and people in developing countries are cut off from a vast array of consumption and production activities relative to a world with 24/7/365 electricity access. Firms from many different sectors that require a continuous supply of electricity cannot enter these markets and existing firms have to constrain their growth or rely on costly diesel

generators (Allcott et al. 2016). Households, rationed off the grid altogether, substitute to costly alternatives like diesel and off-grid solar power, or forego electricity entirely when given these inferior options (Burgess et al. 2019). They are consequently unable to make use of a whole range of life-enhancing appliances. We do not observe the latent demand that firms and people have for continuous, reliable electricity because electricity with these characteristics is not offered.

What is the way out? We offer a taxonomy of reform in four areas: explicit subsidy reform, changing social norms, better technology, and privatization. Many of these policies are complements. They share a longer-run goal of changing the way people think about electricity—that is, their aim is to break the social norm that electricity is a right. They are particularly important because countries or regions of countries that have universal electrification as their ultimate goal will need to employ them so that each additional electricity customer is profitable rather than loss-making.

First, countries could reduce explicit subsidies for electricity, both in size and in scope, while continuing to support the poor. Subsidies on electricity are often enjoyed by consumers across the income distribution, which both makes them regressive and furthers the notion that power is an entitlement. For example, government might instead provide direct benefits to the poorest members of society. If needed for the transition, a well-defined category of poor consumers may receive a "tagged" subsidy payment equal to the subsidies they would have received under current subsidized electricity prices. Indonesia is an example of a country that has moved away from energy subsidies towards direct transfers, though its policy has wavered lately (Burke and Kurniawati 2018).

Second, reforms might seek to reduce theft of electricity and nonpayment of bills. In Bihar, we engaged in a large-scale experiment involving 28 million consumers to enact such a scheme. Under this initiative, the hours of electricity provided by the utility to a feeder were explicitly linked to bill collection rates via a transparent and heavily publicized schedule. This policy targets utility supply. However, losses remain high because we can only target payment by groups of 13,000 people but not individual customers. A similar initiative is underway nationally in Pakistan, allowing utilities there to cut off areas that are the most egregious offenders. In these efforts, it is critical to communicate the benefits of paying for electricity. In Bihar, bill inserts, posters, text messages, and public announcements were used to relay how communities paying more would receive longer hours of electricity. Similarly, in Sao Paulo, utilities held meetings with de facto leaders of slums before introducing billing; in Delhi, one utility hired 800 women from informal settlements to act as community liaisons (Lawaetz 2018).

A related set of reforms provides incentives to distribution company employees who collect electricity payments. In theory, these high-performance incentives both elicit greater collection effort and break the collusion whereby consumers offer electricity bribes to the bill collectors, rather than paying for electricity (Khan, Khwaja, and Olken 2016). We are involved in evaluating an experimentally assigned scheme where utility employees in Bihar move from flat payments to one where they also retain a proportion of revenue from bills collected.

Bill collection may be aided by social trust—when the collectors are your neighbors, it is harder to ignore them. Rural electrification in the United States was achieved largely through rural electrification cooperatives, which were groups of farmers that maintained the grid and collected bills (Lewis and Severnini forthcoming; Kitchens and Fishback 2015). The history of electrification in China and South Korea also involved local engagement with the electricity sector. Initial electrification was mainly funded by communities rather than the national government, and in some cases farmers were hired part time as bill collectors (Aklin et al. 2018; Niez 2010). Rural communities were eventually connected to the national grid in the 2000s, but reported electricity losses remained low, perhaps because of early local buy-in (Bhattacharyya and Ohiare 2012).

A third type of reform relies on technology to make electricity excludable, therefore making it possible to explicitly link payments and supply at the individual level. Smart meters can require payments in advance or allow the utility to cut off household electricity supply remotely. Smart meters have been shown to reduce power consumption in some contexts (Jack and Smith 2015). That said, there remains a need for more evidence from high-theft environments because even the best meter does nothing if a consumer connects themselves directly to the line on the street or can wire around a meter. Better monitoring can also be undercut by bureaucratic collusion, as highlighted in the healthcare literature (Banerjee, Duflo, and Glennerster 2008).

Fourth, why not aim to privatize distribution in the hope that this leads to a market for electricity? A comparison often mentioned here is that the cellphone market in many developing countries is run through private markets. However, the political economy of electricity distribution makes the leap to privatization in many developing country contexts difficult. As long as electricity is perceived as a right by all parties, effective privatization is not feasible (Reddy and Sumithra 1997). The case of Odisha, a poor state neighboring Bihar, is illustrative. The state distribution companies were among the earliest in India to be restructured and privatized, but have continued to suffer some of the highest loss rates in the country for two decades (as high as 34 percent as of 2018) and require continued subsidization (PowerLine 2018).

Where privatization has sufficient public support, it might improve efficiency. For example, Delhi privatized electricity distribution in 2002 and has seen incredibly rapid reductions in losses and improvements of supply—partly through the social engagement and technical reforms recommended above. Even so, power prices have remained a political hot button. In 2015, the Delhi government reintroduced a significant 50 percent power subsidy for all consumers who use less than 400 kilowatt-hours per month (Tongia 2017). With this threshold, over 80 percent of households in the city received the subsidy. In September 2019, the subsidy was made even more generous, with consumption below 200 kilowatt-hours made completely free. Beyond the large direct costs of this policy, it remains to be seen whether such a policy might reintroduce the norm of electricity being a right and affect payment behavior more broadly, including among the middle and upper

classes to whom the subsidy sometimes applies. This possibility underscores the fragility of a high payment equilibria when electricity is still seen as a right.

We conclude with the reminder that 24/7/365 electricity remains out of reach for the majority of people in developing countries. Macro solutions, like privatization of the electricity industry or construction of ever more wires and plants, come into and out of favor, but we believe they are targeting the symptoms, not the cause. High losses and poor quality supply will persist, despite ambitious reforms, so long as electricity is treated as a right.

References

Aklin, Michaël, Patrick Bayer, S.P. Harish, and Johannes Urpelainen. 2018. *Escaping the Energy Poverty Trap: When and How Governments Power the Lives of the Poor.* Cambridge: MIT Press.

Akwagyiram, Alexis, and Paul Carsten. 2018. "Exclusive: Nigerian Energy Sector's Crippling Debts Delay Next Power Plant." London: Reuters.

Allcott, Hunt, Allan Collard-Wexler, and Stephen D. O'Connell. 2016. "How Do Electricity Shortages Affect Industry? Evidence from India." *American Economic Review* 106 (3): 587–624.

Asher, Sam, and Paul Novosad. 2017. "Politics and Local Economic Growth: Evidence from India." *American Economic Journal: Applied Economics* 9 (1): 229–73.

Banerjee, Abhijit V. 1997. "A Theory of Misgovernance." *Quarterly Journal of Economics* 112 (4): 1289–1332.

Banerjee, Abhijit V., Esther Duflo, and Rachel Glennerster. 2008. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System." *Journal of the European Economic Association* 6 (2–3): 487–500.

Baskaran, Thushyanthan, Brian Min, and Yogesh Uppal. 2015. "Election Cycles and Electricity Provision: Evidence from a Quasi-experiment with Indian Special Elections." *Journal of Public Economics* 126: 64–73.

Bhattacharyya, Subhes C., and Sanuse Ohiare. 2012. "The Chinese Electricity Access Model for Rural Electrification: Approach, Experience and Lessons for Others." *Energy Policy* 49 (3): 676–87.

Burgess, Robin, Michael Greenstone, Nicholas Ryan, and Anant Sudarshan. 2019. "Demand for Electricity in a Poor Economy." http://www.lse.ac.uk/economics/Assets/Documents/personal-pages/robin-burgess/demand-for-electricity-in-a-poor-economy.pdf.

Burke, Paul J., and Sandra Kurniawati. 2018. "Electricity Subsidy Reform in Indonesia: Demand-Side Effects on Electricity Use." *Energy Policy* 116: 410–21.

Davis, Lucas W., and Lutz Kilian. 2011. "The Allocative Cost of Price Ceilings in the U.S. Residential Market for Natural Gas." *Journal of Political Economy* 119 (2): 212–41.

Dinkelman, Taryn. 2011. "The Effects of Rural Electrification on Employment: New Evidence from South Africa." *American Economic Review* 101 (7): 3078–3108.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–78.

Engelmeier, Tobias. 2015. "Will India's Power Sector Run Out of Debt?" https://www.linkedin.com/pulse/indias-power-sector-run-out-debt-tobias-engelmeier/.

Fried, Stephie, and David Lagakos. 2017. "Rural Electrification, Migration and Structural Transformation: Evidence from Ethiopia." International Growth Centre Working Paper E-32301-ETH-1.

Fried, Stephie, and David Lagakos. 2019. "Cottage Industry to Factories? The Effects of Electrification on the Macroeconomy." Paper presented at the Econometric Society Winter Meetings, Atlanta, GA.

GhanaWeb. 2018. Power Rationing Imminent as GRIDCo Reels under Huge Debt. https://www.ghanaweb.com/GhanaHomePage/NewsArchive/Power-rationing-imminent-as-GRIDCo-reels-under-huge-debt-683468#.

Government of India Ministry of Power. 2019. *UDAY National Dashboard*. https://www.uday.gov.in/home.php (accessed November 1, 2019).

Jack, B. Kelsey, and Grant Smith. 2015. "Pay as You Go: Prepaid Metering and Electricity Expenditures in South Africa." *American Economic Review* 105 (5): 237–41.

Kassem, Dana. 2018. "Does Electrification Cause Industrial Development? Grid Expansion and Firm Turnover in Indonesia." Universität Bonn Discussion Paper CRC TR 224.

Khan, Adnan Q., Asim I. Khwaja, and Benjamin A. Olken. 2016. "Tax Farming Redux: Experimental Evidence on Performance Pay for Tax Collectors." *Quarterly Journal of Economics* 131 (1): 219–71.

Kitchens, Carl, and Price Fishback. 2015. "Flip the Switch: The Impact of the Rural Electrification Administration 1935–1940." *Journal of Economic History* 75 (4): 1161–95.

Kumi, Ebenezer Nyarko. 2017. *The Electricity Situation in Ghana: Challenges and Opportunities*. Washington, DC: Center for Global Development.

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

Lawaetz, Simone. 2018. *Advancing Power Sectors' Self-Reliance through Electricity System Loss Reduction*. Washington, DC: Energy at US Agency for International Development.

Lee, Kenneth, Edward Miguel, and Catherine Wolfram. 2019. "Experimental Evidence on the Economics of Rural Electrification." https://www.dropbox.com/s/ougxs9jtchzkj3m/REPP-Paper_2019-02-04_Final.pdf?dl=0.

Lewis, Ethan, and Edson Severnini. Forthcoming. "Short- and Long-Run Impacts of Rural Electrification: Evidence from the Historical Rollout of the U.S. Power Grid." *Journal of Development Economics*.

Lipscomb, Molly, A. Mushfiq Mobarak, and Tania Barham. 2013. "Development Effects of Electrification: Evidence from the Topographic Placement of Hydropower Plants in Brazil." *American Economic Journal: Applied Economics* 5 (2): 200–31.

Mahadevan, Meera. 2019. *The Price of Power: Costs of Political Corruption in Indian Electricity*. https://conference.nber.org/conf_papers/f130302.pdf.

McRae, Shaun. 2015. "Infrastructure Quality and the Subsidy Trap." *American Economic Review* 105 (1): 35–66.

Moneke, Niclas. 2019. "Can Big Push Infrastructure Unlock Development? Evidence from Ethiopia." https://niclasmoneke.com/wp-content/uploads/Moneke-JMP-Big_Push_Infrastructure.pdf.

Morales, Evo. 2012. *Manifiesto de la Isla del Sol: Ten Commandments against Capitalism, for Life and Humanity*. La Paz: printed by the author.

Modi, Narendra. 2019. *When Development Ensures Dignity*. https://www.narendramodi.in/te/when-development-ensures-dignity-14-march-2019-544034 (accessed November 1, 2019).

Nigerian Electricity Regulatory Commission. 2019. *ATC&C Losses*. https://nerc.gov.ng/index.php/library/industry-statistics/distribution/119-atc-c-losses.

Niez, Alexandra. 2010. "Comparative Study on Rural Electrification Policies in Emerging Economies." International Energy Agency Information Paper 2010/03.

Pargal, Sheoli, and Sudeshna Ghosh Banerjee. 2014. *More Power to India: The Challenge of Electricity Distribution*. Washington, DC: The World Bank.

PowerLine. 2018. *Looking Up: Utilities Report Overall Improvement in Operational Performance*. https://powerline.net.in/2018/09/09/looking-up-2/ (accessed November 1, 2019).

Prateek, Saumy. 2018. "BERC Sets 4.16/kWh as Average Power Purchase Cost for Bihar DISCOMs." *MERCOM India*, March 23. https://mercomindia.com/berc-power-purchase-cost-discoms/.

PTI. 2018. "Over 70% Uday Bonds to Mature during 2019-27: Icra." *Money Control*, Nov 19. https://www.moneycontrol.com/news/business/over-70-uday-bonds-to-mature-during-2019-27-icra-3197661.html.

Reddy, Amulya K., and Gladys D. Sumithra. 1997. "Karnataka's Power Sector: Some Revelations." *Economic and Political Weekly* 32 (12): 585–600.

Shaukat, Mahvish. 2018. Too Close to Call: Electoral Competition and Politician Behavior in India. https://economics.mit.edu/files/16553.

Sustainable Energy for All (SEFA). 2012. *A Global Action Agenda: Pathways for Concerted Action toward Sustainable Energy for All*. Vienna: Sustainable Energy for All.

Tongia, Rahul. 2017. *Delhi's Household Electricity Subsidies: High and Inefficient*. New Delhi: Brookings India.

Walsh, Mary Williams. 2016. "How Free Electricity Helped Dig $9 Billion Hole in Puerto Rico." *The New York Times*, February 1. https://www.nytimes.com/2016/02/02/business/dealbook/puerto-rico-power-authoritys-debt-is-rooted-in-free-electricity.html.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 170–195

Solo Self-Employment and Alternative Work Arrangements: A Cross-Country Perspective on the Changing Composition of Jobs

Tito Boeri, Giulia Giupponi, Alan B. Krueger, and Stephen Machin

In the last 20 years, most OECD countries experienced a major change in the composition of self-employment. The share of self-employed persons who operate on their own without having dependent workers on their payroll—or *solo self-employment*—increased almost everywhere relative to the other self-employment. This changing nature of self-employment raises a number of relevant issues: is solo self-employment an intermediate status between employment and unemployment? Does it contribute to explaining the strong wage moderation that OECD countries are experiencing even in the presence of low-measured unemployment? Are policies encouraging self-employment as a vehicle for entrepreneurship and job creation ill-suited for these new developments? How do the preferences of the solo self-employed locate along the trade-off between flexible work organization and income insecurity imposed by their working arrangements? Is there a need to extend social protection to these new forms of employment? If so, how is this possible?

Economic theory typically treats self-employment as a labor supply decision. Most of the economic literature on self-employment is focused on

■ *Tito Boeri is Professor of Economics, Bocconi University, Milan, Italy. Giulia Giupponi is Postdoctoral Fellow, Institute for Fiscal Studies, London, United Kingdom. In September 2020, she will be Assistant Professor of Economics, Bocconi University, Milan, Italy. Alan B. Krueger was the James Madison Professor of Political Economy, Princeton University, Princeton, New Jersey, before his death on March 16, 2019. Stephen Machin is Professor of Economics and Director of the Centre for Economic Performance, London School of Economics, London, United Kingdom.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.170.

entrepreneurship (Evans and Jovanovic 1989; Jovanovic 1994; Parker 2004; Lazear 2004; Audretsch, Keilbach, and Lehmann 2006). A partial exception is Levine and Rubinstein (2017), who acknowledge the difference between self-employment in incorporated and in unincorporated enterprises but do not consider the demarcation between solo self-employed and self-employed with employees. The broader theoretical framework used in this literature is a model of occupational choice in which workers make a voluntary choice either to be self-employed or in paid employment, based on factors like their skills and degree of risk aversion. Some workers might prefer greater flexibility in organizing their time or other nonpecuniary benefits of being their own boss (as in Hurst and Pugsley 2011). By treating self-employment as a choice, this framework does not allow for demand-driven determinants of self-employment. For example, it does not allow for employers who are unwilling to offer employment protection to individuals who are de facto dependent workers in their enterprise.

But do self-employed workers agree that they have made an occupational choice that they prefer to conventional dependent employment? Self-employed people without employees do not have the same type of social insurance and job protection that are granted to employees. Some countries have a dual labor market with a substantial number of fixed-term contract holders, but even compared with this group, self-employed individuals do not have any protection even within the contract duration and frequently are not covered by the various forms of social insurance provided to workers with fixed-term contracts.

The purpose of this paper is to shed fresh light on the situation of self-employed workers, with a particular emphasis on solo self-employment, drawing on newly collected survey data investigating the Italian, UK, and US labor markets. In these three countries, we conducted comparable surveys of self-employment, alternative work arrangements, and the gig economy, including questions on demographics, job characteristics, contractual conditions, the need for flexibility, and willingness to pay for social protection. We complement these data with information on macro trends from OECD data and on individual labor market dynamics from the UK and Italian Labor Force Surveys (LFS) and the US Current Population Survey (CPS). This provides a unique international comparison of the changing nature of self-employment in three major economies.

We first consider the data on self-employment with and without workers available from the OECD. Self-employment with employees is falling in most countries, while solo self-employment is rising in nearly half of them. As a consequence, the solo component of self-employment is increasing relative to self-employment with employees almost everywhere. A recurrent theme of this paper is that the solo self-employed differ from the self-employed with employees. We also deal with measurement issues, which are extremely important when dealing with self-employment, and the relationship between self-employment and alternative work arrangements like gig work.

We then turn to our surveys of workers in the United Kingdom, the United States, and Italy to describe how the characteristics of the workers engaged in

solo self-employment compare to self-employed with workers and the reasons why workers engage in these types of jobs. In particular, we investigate the extent to which nonstandard work arrangements satisfy the need for flexibility, or by contrast, whether workers engaged in solo self-employment would prefer to work more hours but are somehow constrained in doing so. There are a number of reasons to suspect that a substantial number of self-employed may not be in search of flexibility. Self-employment contracts frequently hide de facto dependent employment conditions with little, if any, working time flexibility. Thus, even workers valuing higher flexibility may be worse off with lower protection against labor market risk and only slightly more flexibility. Indeed, we present evidence, especially among gig workers, of a bimodal distribution of the degree of job satisfaction, with more or less the same proportion of workers being hourly constrained and being happy about their current hours. This sits well with the recent study of US call center applicants that found that the majority of workers do not value workplace flexibility and have a strong distaste for irregular and short-noticed scheduling (Mas and Pallais 2017).

We then turn to the labor market dynamics of workers to consider the transition patterns in and out of unemployment, regular employment, solo self-employment, and self-employment with employees. Again, strong evidence emerges that solo self-employment and self-employment with workers are two distinguishable labor market statuses, characterized by different transitions from and into unemployment. Moreover, solo self-employment is largely associated with underemployment: that is, these workers would like to work more hours, and they earn less on an hourly basis than their counterparts with employees. The solo self-employed are also more liquidity constrained and more vulnerable to idiosyncratic shocks than the self-employed with workers.

These features of solo self-employment make it a candidate to be considered as part of an overall measure of labor slack. Indeed, we will argue that labor market slack may no longer be captured by unemployment and involuntary part-time figures alone, especially in European labor markets. Even in countries with very low unemployment levels there is now a large "reserve army" in place, including some of the solo self-employed, that potentially undercuts wages of those working in traditional forms of employment.

We also discuss the demand and supply of social protection and the problems to be addressed by reforms that could possibly extend work injury, sickness, old age, and unemployment insurance to these solo self-employment work arrangements. Our surveys indicate that the solo self-employed express a strong demand for social protection and are willing to pay even more than the rate charged to the traditional forms of employment in order to get some social insurance coverage. The key challenge is how to design social protection for self-employed who can readily alter their working status and incomes and how to address the problems of moral hazard and adverse selection that arise. In the conclusion, we offer some policy recommendations and directions for further research.

Self-Employment: Basic Trends and Measurement Issues

Trends

The OECD definition of self-employment refers to "individuals who are the sole owners, or joint owners, of the unincorporated enterprises in which they work, excluding those unincorporated enterprises that are classified as quasi-corporations." In our discussion, we will focus mainly on the separation between self-employed with and without employees. This difference is better understood by survey respondents, and as we will see, it demarcates quite a different employment dynamic.

Table 1 shows trends in self-employment rates (the ratio of the number of self-employed to total employment) in OECD countries. In most countries, self-employment has been declining as a share of total employment, and the strongest declines are observed in those countries that had in 2000 the highest self-employment rates—typically southern European countries. However, cross-country differences in self-employment rates were still sizeable in 2017, ranging from a low 6 percent in Norway to a high 30 percent in Greece. Such a large cross-country variation is a by-product of institutional asymmetries, such as the strictness of employment protection legislation and differences in the structure of employment (namely the relevance of the small business sector, notably in retail trade). As employment protection legislation is declining in most countries as a result of reforms introducing more flexible forms of dependent employment and globalization has brought about an increase in scale economies, the very same factors explaining why some countries had historically high self-employment rates contribute to explaining the fall of the overall share of self-employment.

However, the fall of self-employment is largely concentrated on self-employment with dependent employees, since self-employment without employees has actually been increasing relative to total self-employment in almost all of the OECD countries.

Self-employed workers with and without employees sort into different occupations. We looked at what main occupations of self-employed with and without employees are the three countries in which our main analysis focuses, using the Labor Force Surveys for the United Kingdom and Italy and the Current Population Survey for the United States. Whilst the main occupations for the self-employed with employees are production or retail manager in all three countries, with the addition of medical practitioner in the United Kingdom, the corresponding occupations among the solo self-employed are taxi driver, carpenter, and childminder in the United Kingdom; manager, farmer, and construction laborer in the United States; and shopkeeper, lawyer, and sales agent in Italy.

The occupations that grew the most among the solo self-employed are professional, technical, and personal care occupations in the United Kingdom, transportation and managerial occupations in the United States, and professional and technical occupations in Italy.[1] Multiple data sources have documented the phenomenal growth of gig

[1] The change is computed over the period 2000–2017 for the United Kingdom and Italy and 2014–2017 for the United States.

Table 1
Self-Employed with and without Employees as a Percent of Total Employment

	Self-employment as a share of total employment		*Solo self-employment as a share of self-employment*	
	2000	*2017*	*2000*	*2017*
Australia	19.13	16.74	60.53	63.14
Austria	10.56	10.57	53.03	56.67
Belgium	13.65	13.07	67.11	69.17
Canada	14.96	13.33	64.71	70.22
Czech Republic	14.36	16.14	70.89	81.29
Denmark	8.03	7.36	47.57	59.10
Finland	12.59	11.66	66.40	67.50
France	9.92	10.89	57.16	62.72
Germany	9.69	9.08	49.95	54.85
Greece	31.44	29.37	74.78	75.79
Hungary	14.40	9.66	65.00	53.31
Iceland	16.88	10.79	57.88	65.89
Ireland	16.77	13.35	65.30	68.46
Italy	23.65	20.86	47.06	72.34
Korea	27.73	21.26	75.12	71.87
Latvia	10.20	11.83	59.71	60.86
Netherlands	10.04	15.51	68.23	74.53
New Zealand	19.72	20.03	64.25	66.40
Norway	6.94	5.87	75.50	70.70
Poland	21.83	17.38	82.27	77.45
Portugal	20.43	13.47	69.55	66.30
Slovenia	9.52	11.40	70.48	66.49
Spain	17.76	15.68	68.81	68.69
Sweden	9.87	8.60	60.39	59.77
United Kingdom	11.48	14.06	72.65	84.00
United States	10.63	10.03	73.85	77.07

Source: OECD.
Note: The table reports the number of (1) self-employed (with and without employees) as a percent of total employment and (2) the share of solo self-employed out of total self-employment for various OECD countries in 2000 and 2017.

economy jobs in the passenger transportation industry in the United States since 2013 (Hall and Krueger 2018; Farrell, Greig, and Hamoudi 2019; Abraham et al. 2018).

Some Caveats about Survey Data on Self-Employment

In survey data, workers are often confused about the nature of their employment relationship; for example, two gig workers out of three in the Italian survey report that they have no clue as to their contractual status. Furthermore, not all surveys have information on the limited liability nature of the business or its legal identity, which prevents classifying the enterprise either as incorporated or unincorporated. For these reasons, the statistical definition of self-employment is often implemented by considering the size of the enterprise. If the firm is relatively small, the worker is classified as a "self-employed person with dependent employees;" if the firm is large,

the worker is classified as an "entrepreneur." This proxy has obvious shortcomings, importantly including the neglect of the age of the firm. Many incorporated business start-ups begin relatively small and then grow.

If the focus is on self-employed people without dependent employees, another issue arises related to the border between self-employment and dependent employment status. Workers classified as self-employed with apparent autonomy over working hours may have a unique client. Indeed, many services offered formally as self-employment activities may not be different from activities carried out by the employees. For this reason, a number of self-employed freelancers, homeworkers, and commission salespersons can be viewed as belonging to an intermediate category between dependent employment and self-employment. So-called gig workers, like those involved in food delivery, sometimes have a status of employee with flexible hours and in other cases are self-employed workers, depending on the choices made by the firm.

Finally, survey data may underestimate the extent of self-employment as they often do not accurately track multiple job holdings. In the United States, for instance, there is evidence of a growing number of self-employed people who are registered in administrative data, but do not show up in survey data. In order to understand the sources of these discrepancies, Abraham et al. (forthcoming) link individual survey data and administrative records. They find that the amount of undocumented self-employment (in Current Population Survey data but not in administrative records) has been relatively stable, while there has been a notable increase in self-employment activity registered by the Internal Revenue Service (IRS) but not by CPS data, and conclude that the latter discrepancy is due—in equal proportions—to underreporting, multiple job holdings, and employment misclassification in the Current Population Survey. In Italy, multiple job holdings seem to be the key factor: registered (at social security) self-employment positions are almost 30 percent of the total registered positions, while the share of self-employed persons in total employment is about 23 percent according to both Labor Force Survey and administrative data. Similarly, in the United Kingdom, Labor Force Survey and administrative tax data converge in reporting a self-employment rate of the order of 12–13 percent, but one self-employed out of four has multiple jobs.

In light of these measurement issues, in this paper we focus mainly on the composition of self-employment, notably on separation between self-employed with and without employees. This difference is better understood by the respondents, and it actually demarcates quite a different employment dynamic, as we have already seen. Furthermore, the most relevant issues nowadays relate to self-employment without employees. Are these solo self-employed activities preferred to dependent employment because they allow for more flexibility in organizing working time? Are the nonpecuniary benefits of being "her own boss" (Hurst and Pugsley 2011) prevailing over the security offered by standard dependent employment contracts? Or is this a choice imposed by the employers willing to share with the worker the enterprise risk by not offering employment protection to persons who are de facto dependent workers of their enterprise?

This issue has been largely overlooked by the academic literature on self-employment. The latter focused almost entirely on self-employment as entrepreneurship—adopting a theoretical framework of voluntary sorting into self-employment by individuals—and devoted much less attention to demand-driven determinants of self-employment (Evans and Jovanovic 1989; Jovanovic 1994; Parker 2004; Lazear 2004; Audretsch, Keilbach, and Lehmann 2006).

Alternative Work Arrangements and Self-Employment

A body of previous work has looked at alternative work arrangements in specific countries, without devoting particular attention to solo self-employment. For example, Katz and Krueger (2018) document a large increase in the percentage of US workers engaged as independent contractors, on-call workers, temporary help agency workers, and contract company workers in the last decade. In a follow-up reconciliation across different data sources, Katz and Krueger (2019) conclude that there has been an upward trend in alternative forms of employment in the US labor market, but also emphasize the difficulty of tracking down workers engaged in these new forms of work in commonly used data sources. Other recent work emphasizes the difficulties of identifying alternative work arrangements in US data sources and the blurred boundaries of employment categories that the new forms of work are generating (for example, see Abraham and Amaya 2019; Abraham et al. forthcoming; Jackson, Looney, and Ramnath 2017; and Spreitzer, Cameron, and Garrett 2017).

Similar patterns were found by Datta, Giupponi, and Machin (forthcoming) in countries like the United Kingdom, where the percentage of the workforce that is self-employed without dependent workers and the share of workers on "zero hours contracts" (who agree to be available for work when required, with no guaranteed hours or times of work) have been increasing over time. There is also some US-based evidence that unemployment is predictive of the probability of transitioning to a nonstandard job (Katz and Krueger 2017), but little is known about the types of labor market transitions those workers on solo self-employment experience. Some studies on measures of nonstandard work (OECD 2015, 2018) and on wage moderation (Bell and Blanchflower, forthcoming) do acknowledge the difference between solo self-employment and the total stock of the self-employed, but making such a distinction remains more the exception than the rule.

Two Faces of Self-Employment

In order to better understand the nature of self-employed workers, we designed comparable online surveys of self-employment and alternative work arrangements for Italy, the United Kingdom, and the United States. For the UK labor market, the LSE-CEP Survey of Alternative Work Arrangements is a survey of 20,000 individuals carried out in February 2018. For the US labor market, the Princeton Self-Employment Survey is a survey of over 10,000 individuals conducted in April

2017. For the Italian labor market, the fRDB Survey of Independent Workers is a survey of 15,000 individuals conducted in May 2018. The survey questionnaires are reproduced in online Appendices C (UK survey), D (US survey), and E (Italian survey).

The surveys, run on online platforms, were designed to be representative of the working-age population. The UK survey was based on a representative sample. For the Italian and US surveys, representativeness is achieved using survey weights from the survey provider and from the 2011–2015 American Community Survey, respectively. To assess the representativeness of the survey samples, we compared them to the UK Labor Force Survey, the US Current Population Survey, and the Italian Labor Force Survey. There is a healthy mixture of representativeness across gender, age, and employment status across the three online surveys. As for educational attainment, the distribution in the online surveys and national surveys do not match well, though this is partly due to difficulties in fully homogenizing educational attainment variables across countries and data sources.[2] In spite of the overall good representativeness, there remain concerns related to self-selection in online surveys and to the fact that such self-selection may differ across countries.

The survey questions investigate previously untapped areas of the labor market, collecting novel information on the characteristics and employment conditions of self-employed workers and offering a unique international comparison of working arrangements in the three major economies.

Self-Employment in the Survey Data

In this section, we focus on respondents who identify themselves as primarily self-employed, and we emphasize the distinction between self-employed with and without employees. Our surveys also investigate gig economy workers, which we will discuss in the next section. Table 2 presents descriptive statistics for self-employed workers in the three countries, distinguishing between self-employed with employees and without employees (own account or solo self-employed). Whilst self-employed workers as a group are predominantly male, the proportion of females is consistently higher among the solo self-employed. Similarly, the solo self-employed tend to be slightly older than the self-employed with employees in all countries. The distribution of educational qualifications is roughly similar across the two groups.

Solo self-employed individuals have mean and median hourly earnings that are consistently lower than those of self-employed with employees across the three countries, as shown in Table 2. A similar pattern is found when looking at weekly hours worked. The solo self-employed work on average eight fewer hours per week than the self-employed with employees. Solo self-employed work fewer hours also in comparison to traditional full-time employees, who work approximately 40 hours per week on average. Moreover, solo self-employed are characterized by a much larger incidence of part-time work, with 40 to 50 percent of solo self-employed

[2]For descriptive statistics about the online survey samples and their representativeness, see Table A1 in online Appendix A.

Table 2
Summary Statistics of Self-Employed Workers

	United Kingdom		*United States*		*Italy*	
	Solo	*With employees*	*Solo*	*With employees*	*Solo*	*With employees*
Female	0.44	0.36	0.41	0.21	0.40	0.37
Age	44.81	42.75	47.01	44.88	42.28	41.11
Age 18–24	0.08	0.07	0.03	0.03	0.06	0.08
Age 25–34	0.15	0.21	0.12	0.17	0.21	0.20
Age 35–44	0.22	0.24	0.19	0.27	0.28	0.32
Age 45–54	0.28	0.26	0.41	0.26	0.29	0.25
Age 55–65	0.26	0.22	0.25	0.27	0.16	0.14
Less than high school	0.14	0.12	0.05	0.10	0.01	0.00
High school	0.32	0.34	0.54	0.47	0.27	0.23
Vocational training	0.15	0.09	0.06	0.04	0.29	0.34
Bachelor	0.27	0.25	0.21	0.24	0.13	0.15
Advanced degree	0.11	0.20	0.15	0.15	0.30	0.28
Hourly wage	36.82	52.49	46.71	65.55	60.48	87.64
Hourly wage (median)	11.00	18.00	22.00	25.00	40.00	53.33
Weekly hours	32.26	41.16	36.03	43.67	34.78	42.53
Weekly hours (median)	31.50	40.00	35.00	42.00	40.00	40.00
Proportion working part time (<35 hours per week)	0.52	0.26	0.46	0.18	0.41	0.19
Proportion working part time for economic reasons (<35 hours per week)	0.18	0.05	0.18	0.03	0.12	0.06
Proportion working as traditional employee	0.07	0.20	0.11	0.43		
Total weekly hours (including traditional employment)	33.69	47.46	38.38	57.38		
Number of observations	1,633	228	1,014	299	2,037	367

Source: LSE-CEP Survey, Princeton Self-Employment Survey, fRDB Survey.
Note: The table reports the mean of a set of variables for the samples of self-employed respondents to the online surveys, distinguishing between solo self-employed and self-employed with employees.

working less than 35 hours per week—the corresponding figure for self-employed with employees ranging from 18–19 percent in the United States and Italy to 26 percent in the United Kingdom.

The solo self-employed often state that they are underemployed for economic reasons: 12 percent in Italy and 18 percent in the United Kingdom and the United States declare that they work part-time due to slack business conditions, the inability to find full-time work, or due to seasonal work. Strikingly, the corresponding figure for self-employed with employees is only 3–6 percent. This evidence is consistent with the notion that the solo self-employed face constraints on how many hours they can work due to an unavailability of additional work; indeed, approximately one-third of the solo self-employed would like to work more hours per week (as shown in Table 3). While many of the self-employed with employees would also like

Table 3

Desired Hours, Job Satisfaction, Liquidity Constraints, and Economic Dependency

	United Kingdom		*United States*		*Italy*	
	Solo	*With employees*	*Solo*	*With employees*	*Solo*	*With employees*
A: Desired hours						
More hours	0.27	0.22	0.34	0.30	0.30	0.16
Fewer hours	0.19	0.23	0.19	0.25	0.26	0.44
Satisfied	0.54	0.55	0.47	0.45	0.44	0.40
B: Job satisfaction						
Very satisfied	0.39	0.64			0.15	0.31
Satisfied	0.41	0.29			0.42	0.47
Neutral	0.14	0.05			0.30	0.20
Dissatisfied	0.05	0.00			0.11	0.02
Very dissatisfied	0.01	0.02			0.02	0.00
C: Liquidity constraints						
Able to pay	0.59	0.75	0.65	0.78	0.64	0.84
Pay by borrowing or selling	0.20	0.16	0.22	0.16	0.21	0.12
Unable to pay	0.21	0.10	0.13	0.05	0.14	0.04
D: Number of different clients in 2017						
1					0.16	0.03
2–5					0.24	0.14
6–15					0.20	0.15
16–50					0.20	0.23
More than 50					0.20	0.45
Number of observations	1,633	228	1,014	299	2,037	367

Source: LSE-CEP Survey, Princeton Self-Employment Survey, fRDB Survey.
Note: Panel A reports the distribution of responses to the question: "Would you have preferred to work more or fewer hours last week in self-employment at that wage rate? Or were you satisfied with the number of hours you worked?" Panel B reports answers to the question: "How satisfied are you with working as a self-employed?" Panel C reports answers to the question: "Suppose that you have an emergency expense that costs 500,00 pounds/400,00 dollars/500,00 euros. Based on your current financial situation, how would you pay for this expense? If you would use more than one method to cover this expense, please select all that apply." Responses are grouped into the three categories reported in the table. Panel D shows the distribution of responses to the question: "How many different customers/clients did you work for in 2017?" Answers are reported separately for solo self-employed and self-employed with employees.

to work more hours, the fraction that wants more hours is always 5 to 15 percentage points lower in this category.

Some self-employed individuals may increase their hours and income via multiple job holdings, thus creating overlap between self-employment and traditional employment. Table 2 presents some information on the extent of this overlap with the UK and US surveys. The fraction working as traditional employees is lower among the solo self-employed in both countries (7 versus 20 percent in the United

Kingdom and 11 versus 43 percent in the United States). This interesting difference could indicate that there are fewer, or worse, outside options for the solo self-employed. However, even when taking into account the total number of hours worked in both employment types, a substantial hour differential remains between the self-employed with and without employees.

Across industries, construction and retail stand out as the main industries of self-employment. There do not seem to be substantial differences in the distributions between solo self-employed and self-employed with employees across industries, with the exception of accommodation and food service activities (predominantly with employees); human health and social work activities (predominantly solo); and arts, entertainment, and recreation (predominantly solo). Detailed survey results about the characteristics of the self-employed across the three countries are reported in online Appendix A.[3]

When asked about their degree of satisfaction with self-employed work (in the Italian and UK surveys), respondents turn out to be overall satisfied with their working arrangements, although the solo self-employed display consistently lower degrees of job satisfaction, as shown in Table 3. The degree of flexibility that self-employed work offers seems likely to be the main driver of relatively high levels of satisfaction. The UK survey asked respondents what their main reason is for being engaged in self-employment. Flexibility is by far the most important reason for both groups, followed by the possibility to work from home for the solo self-employed and better pay for those with employees (as shown in Figure A4 in online Appendix A). Importantly, around 12 percent of self-employed report that they took this job because it was the only available option, reflecting that the lack of outside options is also a non-negligible factor.

Underemployment and a lack of outside options may have important consequences for the liquidity constraints of the individual workers. In all three surveys, we ask respondents how they would pay for an unexpected expense of 500 euros (Italy), 500 pounds (United Kingdom), or 400 dollars (United States). Results reported in Table 3 highlight a striking difference between the two groups of self-employed, with the solo self-employed being substantially more liquidity constrained. Across the three countries, approximately two-thirds of the solo self-employed would be able to pay, while the remainder would be evenly split between those who would borrow or sell something and those who would be unable to pay. The same figures for self-employed with employees show that approximately 80 percent would be able to pay for the expense and only very few would be unable to do so.

[3]In online Appendix A, Table A2 reports the industry distribution of self-employed workers in the three countries. Figure A1 reports the empirical distribution of hourly wages for self-employed with and without employees in the three countries. Figure A2 reports the empirical distribution of weekly hours for self-employed with and without employees in the three countries. Figure A3 reports evidence on the reasons why UK respondents are unable to work more hours and why they would like to work fewer hours. Table A3 shows summary statistics on weekly hours for full-time employees based on UK Labor Force Survey, US Current Population Survey, and Italian Labor Force Survey data.

Another dimension that may affect the economic insecurity of the individual worker is the degree of de facto economic dependency from a single client or contractor, a situation in which a self-employed worker is bound to face a higher risk of insecurity in response to idiosyncratic shocks affecting that main client or contractor. In the Italian survey, we asked the number of different clients for which the individual worked in the previous year. For the solo self-employed, the distribution of the number of clients is rather uniform across the different bins, with 16 percent of the sample having only one client. For the self-employed with employees, the latter figure drops to 3 percent and increasingly larger fractions of respondents engage with larger numbers of clients (as shown in Table 3). However, when we asked what share of their total revenue originates from their main client, approximately 20 percent of both solo self-employed and self-employed with employees answered that they are economically dependent on their main client for more than 50 percent of their revenue. This pattern suggests that the degree of economic dependency from a single entity is overall limited, yet with pockets of solo self-employed that face a very high risk of economic insecurity.

It is worth noting that the survey results illustrated so far display substantial uniformity across the three countries. In light of the fact that the countries are characterized by very different labor market institutions, such uniformity lends support to the hypothesis that the duality of self-employment is unlikely to stem from institutional factors, but is rather due to common and pervasive technology, labor demand, or labor supply factors affecting the demand for labor. We document that labor supply factors—such as the preferences for flexibility or, as we will show below, for social protection—do not seem to differ substantially between self-employed with and without employees.

A Focus on "Gig Workers"

Gig economy workers epitomize a shift away from traditional employment toward independent contract work and the trade-off between greater job flexibility and economic insecurity. In our three surveys, we investigate the nature of gig economy workers, though with the caveat that the survey modules on gig economy work are not fully comparable in their definitions and scope across countries. In the UK and US surveys, gig economy workers are considered as a subgroup of primarily self-employed workers and are only surveyed in a limited way. In the Italian survey, the number of questions asked is larger, and a more appropriate and encompassing definition is used, which includes individuals who are (1) primarily gig workers or (2) primarily self-employed or traditional employees and secondarily gig workers.[4]

[4]In the UK Survey, gig workers are defined as a subsample of primarily self-employed workers who answer positively to question Q28 in online Appendix C. In the US Survey, gig workers are defined as a subsample of primarily self-employed workers who answer positively to question Q4 in online Appendix D. In the Italian Survey, gig workers are defined as respondents who answer positively to question SC1 in online Appendix E. Gig work can be their primary or secondary job (in which case, they may be either traditional employees or self-employed in their primary job).

For this reason, we will mainly focus on the Italian survey results and provide comparisons with other countries when suitable.

Consistent with other estimates of the size of the gig economy (Harris and Krueger 2015; Farrell, Greig, and Hamoudi 2019), gig workers make up a small fraction of total respondents in Italy (4 percent) and a limited portion of those who work primarily as self-employed: 5 percent in Italy, 7 percent in the United Kingdom, and 14 percent in the United States. Gig work is characterized by strikingly low hourly wages and weekly hours: 7 euros per hour and 5 hours per week at the median in Italy.

It turns out that gig work is indeed characterized by a high degree of flexibility, since two-thirds of workers can choose freely when to work and almost 80 percent where to work. Such flexibility can be especially valuable in that it offers a self-insurance mechanism in response to income shocks. Consistently with work by Koustas (2018) on ridesharing in the US economy, our survey results indicate that gig work is used to buffer temporary shocks or top-up income by 80 percent of gig workers, but is the only source of income for only 16 percent of them. Compared to the solo self-employed, Italian gig workers appear slightly, though not substantially, more liquidity constrained. However, when compared with the same result for the self-employed, the fraction of gig workers that is hourly constrained is—remarkably—almost 15 percentage points (or 50 percent) higher. Detailed results on gig workers are reported in online Appendix B.[5]

One takeaway from these survey responses of gig workers is that policies which seek to regulate alternative work arrangements by limiting their flexibility may not be desirable, in that they may well harm individuals for whom their gig jobs are usefully used as smoothing devices. From a policy standpoint, concern should be less about the flexibility that gig economy jobs offer and more about poor career development prospects, lack of wage progression, excess uninsured income volatility—especially for those who perform gig work as their main job—and exposure to longevity risk in the presence of low savings rates and limited social protection.

Labor Market Transitions and Wage Moderation

In discussions about the new forms of self-employment and gig work, one prominent recurring question is whether they are forms of employment held by

[5] In online Appendix B, Table B1 reports summary statistics for a set of characteristics of gig economy workers. Figures B1 and B2 show the distributions of hourly wages and weekly hours for gig economy workers. The distributions are spectacularly right-skewed, indicating that gig work is a predominantly short-hour, low-pay activity. Table B2 reports survey responses to questions related to desired hours, job satisfaction, the reasons for working in the gig economy, job flexibility, and liquidity constraints. The results highlight a stark dichotomy between those for whom such short hours are a constraint (that is, who would like to work more hours) and those who are instead happy with their current hours. They also show that gig workers are much less satisfied with their working arrangements than self-employed workers.

individuals because they are the only option they have available, while the individuals would prefer something else, or whether such employment relationships are chosen because the worker places a high value on factors like greater flexibility and independence at work. This section offers empirical evidence on this from two standpoints. The first looks at labor market transitions to ascertain the extent to which individuals are more likely to move in or out of these work arrangements from different prior states of labor market participation (principally from "regular" employment, self-employment, or unemployment). The second looks at whether these new forms of employment are placing downward pressure on wages, which would follow if the individuals employed in them are more likely to be taking these forms of work in the absence of other employment opportunities.

Labor Market Transitions

This section offers evidence on the labor market transitions of individuals in the United Kingdom, the United States, and Italy for transitions taking place between 2016 and 2017. Since the analysis of labor market transitions requires the use of longitudinal survey data at the individual level, we turn to nationally representative longitudinal surveys: the UK and Italy evidence comes from their respective quarterly Labor Force Surveys, the structure of which permits annual transitions between (in this case) 2016 and 2017 to be studied; the US evidence comes from the Current Population Survey, which has a longitudinal setup such that individuals are in the survey for four months, they then drop out for eight months, but return in the same four months in the subsequent year. This too permits the study of transitions between 2016 and 2017.

Table 4 reports the unconditional probabilities of transitioning from a given labor market state in 2016 into different labor market states in 2017 for each of the three countries. The sample is a balanced panel of individuals aged 18–65 in 2016 and in the labor force in both 2016 and 2017. As the tables show, workers in a certain state of the labor market in 2016 are likely to remain in that state in 2017, as one can see by reading the diagonal entries.[6]

But our focus here is on the minority who do switch work states, and a highly consistent pattern of results emerges across the three countries. First, individuals are significantly more likely to enter solo self-employment from unemployment than from traditional employment. The increasingly important group taking solo self-employed positions are indeed mostly coming from unemployment, and this squares up well with the earlier survey results showing that low wages and poor labor market protection are a feature of these jobs.

Second, the patterns of self-employment with employees are different. This group is the least likely to keep the same job status from year to year. In the UK and US data, those changing away from self-employment with employees are roughly equally likely to end up as regular employees or solo self-employed; in

[6]On state dependence in labor market states more generally, see, inter alia, Heckman (1981), Hyslop (1999), or in the case of self-employment Henley (2004).

Table 4
Transition Matrices

	Status in t				
Status in t − 1	*Unemployed*	*Employee*	*Solo SE*	*SE with empl.*	*Total*
A: UK LFS					
Unemployed	44.20	50.02	5.79	0.00	100
Employee	1.28	96.54	1.87	0.31	100
Solo SE	1.05	10.41	85.68	2.87	100
SE with employees	0.00	16.42	20.28	63.31	100
Total	2.81	84.13	11.44	1.63	100
B: US Current Population Survey					
Unemployed	26.41	69.08	4.18	0.33	100
Employee	2.03	95.17	2.26	0.54	100
Solo SE	0.83	30.62	60.53	8.02	100
SE with employees	0.27	25.39	22.78	51.56	100
Total	2.63	87.97	7.11	2.30	100
C: IT LFS					
Unemployed	64.01	32.27	3.31	0.42	100
Employee	2.46	96.96	0.44	0.13	100
Solo SE	1.78	2.91	86.77	8.55	100
SE with employees	0.69	1.79	19.23	78.29	100
Total	7.02	73.81	13.36	5.81	100

Source: UK Labor Force Survey, Current Population Survey, Italy Labor Force Survey.
Note: The table reports transition matrices of the unconditional probability of transitioning from labor market status j in year $t-1$ into labor market status k in year t. The samples are balanced panels of individuals aged 18–65 in year $t-1$ and in the labor force in both year t and $t-1$. Panel A uses the longitudinal version of the UK Labor Force Survey for years 2016/2017 (all quarters). Panel B uses the longitudinal version of the Current Population Survey for years 2016/2017 (all months). Panel C uses the longitudinal version of the Italy Labor Force Survey for years 2016/2017 (all quarters).

Italy, by contrast, very few of the self-employed with employees switch to regular employee status. In Italy, the incidence of self-employment with a small number of employees is higher, possibly indicating that some of their jobs may be somewhat less entrepreneurial in nature and could partly reflect opportunities for those unable to secure "regular" employment. More generally, it is possible that some of the solo self-employed are previously self-employed with employees whose business activity has declined.

Third, the self-employed are less likely to transition into unemployment, compared to traditional employees. In addition, the solo self-employed are always more likely than self-employed with employees to transition into unemployment. Thus, solo self-employment emerges as an intermediate state between traditional employment and self-employment with employees.

Fourth, there is some indication that self-employment without employees may be the initial stage of a future entrepreneurial activity with employees: in this respect, the self-employed without employees are more likely than the unemployed

or the employees to become self-employed with employees. The transition probabilities, though, are rather small, suggesting that this is a limited phenomenon. [7]

Overall, these findings accord well with discussions of how there has been an expansion of a less clearly defined hinterland in the labor market between employment and self-employment, where these independent contractors undertake their work.

Wage Moderation

If the new forms of work are in part reflecting that people moving into these jobs do not have many alternatives, have poor outside opportunities, and are underemployed in that they would like to work more hours, then this may have ramifications for overall wage growth. This argument has been made by Bell and Blanchflower (forthcoming), who argue that the official unemployment rate does not these days measure labor market slack very well. The unemployment rate thus underestimates the number of individuals who would like conventional employment but cannot get it and instead end up in self-employment, perhaps of the gig work variety. In this paper, we place more structure to the argument by considering underemployment, but also thinking that there is more slack because of the new forms of employment—both solo self-employment and gig work—that are present in today's labor market and were not there 10 or 15 years ago.

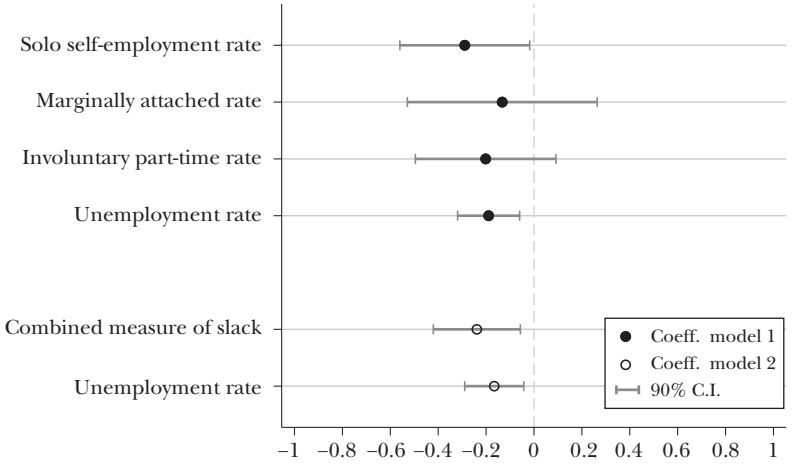
Bell and Blanchflower (forthcoming) provide empirical support for a wider definition of labor market slack by showing that inclusion of underemployment variables in traditional wage curves (for example, à la Blanchflower and Oswald 1995a,b) adds explanatory power over and above the conventionally considered unemployment rate. They show an extra negative effect on real wages resulting from underemployment in their wage curves estimated for Europe and the United States.

In our own work, we consider some cross-country panel regressions of OECD countries, using hourly wage growth as the dependent variable. A model in the style of Hong et al. (2018) uses lagged inflation, productivity growth, the unemployment rate, and the change in the unemployment rate as explanatory variables. However, we find that when a variable for solo self-employment is added to the explanatory variables, it has an additional statistically significant effect in line with the notion that it too reflects some degree of slack in the labor market. In particular, there is evidence that a higher share of solo self-employed is associated with lower wage growth.[8] This can be seen from Figure 1, which provides a graphical representation of the estimated effect of measures of labor market slack on hourly wage growth.

[7] A regression analysis of this data shows that the patterns mentioned in the text are statistically significant at conventional levels. Table A4 in online Appendix A presents a series of regressions for each country, using different market outcomes in 2017 as the dependent variable and then using labor market status in 2016 as the key explanatory variable. Shifts from one labor market status to another are estimated conditional on the lagged value of the dependent variable on the right-hand side (to capture state dependence) as well as a set of control variables for factors like gender, age, and education.

[8] See Table A5 and Figure A5 in online Appendix A for further details.

Figure 1
Estimated Effect of Measures of Labor Market Slack on Hourly Wage Growth



Source: OECD.
Note: The graph reports the point estimates and confidence intervals of a set of coefficients for an "augmented" wage curve estimated on a panel of OECD countries. The circles show the estimated effect of the variables reported on the *y*-axis on hourly wage growth at the country level. The solo self-employment rate is computed as the share of solo self-employed over total employment, the marginally attached rate as the share of marginally attached over total employment, and the involuntary part-time rate as the share of involuntary part-timers over total employment. The "combined measure of slack" is the sum of involuntary part-timers, marginally attached, and solo self-employed (all as a share of total employment). Details on the regression are reported in Table A5 in online Appendix A. The black circles correspond to estimates reported in column 5 of Table A5, the hollow circles to those in column 6. Each model's coefficients are jointly estimated and conditional on lagged inflation, the change in the unemployment rate, a moving average of labor productivity growth, country fixed effects, and year fixed effects.

Thus, there is some evidence that wage growth does seem to have been dampened by the diffusion of new forms of self-employment. This is supportive of the idea that some of these jobs are marginal, in the sense that they are being taken in some cases by workers with not much alternative, and so are inducing more labor market slack than the regular unemployment rate measures. Of course, many of these solo self-employment jobs are also characterized by poor provision of nonwage benefits through social protection, and we turn to the issue that frames the desirability or otherwise of the whole job package in the next section.

Social Insurance

The existence of solo self-employment jobs, gig work, and other forms of alternative work arrangements raises some difficult questions for social insurance. In most countries with a formalized welfare state, those in dependent employment—at

least those with larger formal employers—are covered by a range of employment rights including minimum wages, statutory holiday and sick pay, old age and survivor pensions, as well as parental leave. The self-employed are not always eligible for these nonwage benefits; indeed, this fact is sometimes put forward as a justification for the differential tax treatment of self-employed workers (OECD 2019). For instance, about one-third of OECD countries do not have an unemployment benefit system for self-employed workers. Maternity benefits are everywhere less generous for the self-employed. Sickness, invalidity, and injury benefits in most of the cases involve an insurance franchise, which is not envisaged for employees. Pensions also offer a lower coverage and are often less generous than for dependent employment. The rationale for this lower generosity and coverage of social insurance is that moral hazard problems are more serious in the case of self-employment. Yet, if self-employment gets closer and closer to a dependent employment status, this justification is no longer applicable.

The question of who is or is not an "employee" and thus eligible for full social insurance has been controversial. The 2017 Taylor Review of Modern Work Practices in the United Kingdom emphasized this issue, especially in the context of gig workers (Taylor 2017). In several high-profile court cases, self-employed individuals legally challenged the companies that classified them as self-employed independent contractors, rather than as employees: for example, such cases have been brought to court by currently self-employed individuals working for Uber and Pimlico Plumbers in the United Kingdom, by Foodora riders in Italy, by Take Eat Easy deliverymen in France, and by Dynamex delivery drivers in California.

Demand for Social Protection

Given the income insecurity and lack of access to employment rights that self-employed workers face, it is not surprising that they express a strong demand for social insurance.

In the UK and US surveys, we elicited opinions of the self-employed about the proposal to establish "Shared Security Accounts," whereby all workers would have social insurance and social security coverage funded through contributions paid in by their employers, contractors, or online platforms (Hanauer and Rolf 2015, Krueger 2018). In particular, we asked the following: "Policymakers have been discussing the idea of creating a fund to help self-employed workers obtain work-related benefits, such as health insurance and retirement savings, that they would be able to receive regardless of where they worked, and they could take with them if they changed jobs. Do you think this is a good idea?" The vast majority (approximately 80 percent) in the two countries and self-employment groups think it is a good idea.[9] There does not appear to be any substantial heterogeneity between self-employed with and without employees—the latter being, if anything, slightly more in favor of creating a fund. Of course, this question does not specify

[9] See Table A6 in online Appendix A for details.

Table 5
Benefit Ranked First

	United Kingdom		*United States*		*Italy*	
	Solo	*With employees*	*Solo*	*With employees*	*Solo*	*With employees*
Retirement savings	0.40	0.46	0.16	0.15	0.42	0.34
Unemployment insurance	0.12	0.09	0.07	0.05	0.15	0.22
Paid sick leave	0.22	0.18	0.03	0.03	0.10	0.08
Health insurance	0.06	0.07	0.52	0.44		
Life insurance	0.05	0.07	0.06	0.10		
Worker compensation insurance	0.05	0.03	0.08	0.13	0.09	0.11
Paid family leave	0.05	0.05	0.04	0.03		
Disability insurance	0.05	0.05	0.04	0.07		
Maternity leave					0.10	0.12
Family allowance					0.14	0.13
Number of observations	1,633	228	1,014	299	2,037	367

Source: LSE-CEP Survey, Princeton Self-Employment Survey, fRDB Survey.
Note: The table show the distribution of responses to the question: "If the government were to help you obtain benefits, which one would be most desirable to you personally?" Answers are reported separately for solo self-employed and self-employed with employees in the United Kingdom, the United States, and Italy.

who would pay for it, and as we discuss in the next section, designing social insurance for self-employed workers raises some tough questions.

We also asked survey respondents in the three countries to rank a list of possible benefits from the most to the least desirable (randomly changing the order in which the benefits were listed across respondents). Table 5 reports the result. A social program for retirement savings was by far the top choice among the self-employed in Italy (35–40 percent) and the United Kingdom (40–45 percent), while health insurance was the most preferred social program for the US self-employed (45–50 percent). Interestingly, no substantial differences emerge between self-employed with and without employees, indicating that the two are rather homogeneous in their preferences for social protection. Also, gig workers as distinguished in our Italian survey seem to have preferences over social protection that are very similar to those of solo self-employed individuals.

In the US survey, we investigated in more depth the extent to which the solo self-employed and the self-employed with employees already had health insurance or a tax-deferred retirement account. For US workers, the solo self-employed are somewhat less likely to have health insurance coverage than self-employed with employees (76 versus 86 percent) and much less likely to take advantage of a tax-deferred retirement savings account (28 versus 60 percent). The solo self-employed are also substantially less likely to use a third party to assist them in gaining benefit coverage (7 versus 34 percent) and are less willing to provide

tax data to a third party to gain such assistance (41 versus 63 percent).[10] This differential in health insurance coverage—which takes on added importance if compared to health coverage rates close to 90 percent for traditional employees (Jackson, Looney, and Ramnath 2017)—is suggestive of unmet demand for social protection.

Potential Supply of Social Insurance

It is difficult to design social insurance schemes for self-employed workers. For example, it is not clear who should pay the employers' contributions. If a solo self-employed person works for a single client, then presumably the client could be made liable for these contributions. However, rules that apply only to those with a single client will encourage them to hire workers only on a part-time or temporary basis, and coordinating cost-sharing rules across multiple clients is complex.

One option is to use platforms to coordinate across employers. The Italian social security administration (INPS) takes this approach in covering some gig workers by requiring their employers to register to the online platform managed by INPS and to pay the worker in advance together with the social security contributions. This system also protects the self-employed against the risk of not being paid by their clients, which can be substantial. In the US survey, we asked the respondents whether in the last year they had at least one incident in which they were not paid on time or not paid in full for a job or project that they completed. We find that 36.1 percent were not paid on time (the figure being 31.8 percent for solo self-employed and 51.3 percent for self-employed with employees). The German artists' insurance—a special scheme that offers artists and writers insurance at a subsidized rate involving mandatory membership for low-earning artists—also charges the final customers for the contributions to social security (Tobsch and Eichhorst 2018).

However, charging employers for social insurance in the presence of an elastic demand for labor means that the incidence of these costs will fall onto self-employed workers in terms of lower prices for their services. In the case of pensions and many other social security contributions that are earnings-related, this makes social insurance into a forced savings plan with a substantial cost borne by the self-employed.

An alternative would be to pay social security contributions for self-employed workers out of general government revenues. However, this approach will raise issues of fairness vis-à-vis other categories of workers, notably low-wage employees. More importantly, moral hazard can make a government-paid system extremely expensive. For example, the self-employed have some control over the timing of their employment and payments, which can complicate the assessment of their eligibility for social insurance. It is precisely for this reason that most countries do not have unemployment benefit schemes covering self-employed workers. A partial exception is provided by the Italian DISCOLL, a program introduced in

[10] For details of the response to these questions, see Table A7 in online Appendix A.

2015 targeting self-employed persons without employees who contributed to the social security system as independent collaborators and who then lost their job. The maximum duration of this benefit is one-half of the months of contribution since the beginning of the year predating the job loss for a maximum of six months. The initial replacement rate is initially 75 percent with a cap at 1,300 euros and declining after the third month. This scheme has provided supplementary income to about 22 percent of the eligible population in the first year and 40 percent in the second year (INPS 2018). There is no evidence that this led to increasing flows from independent collaborators to unemployment.

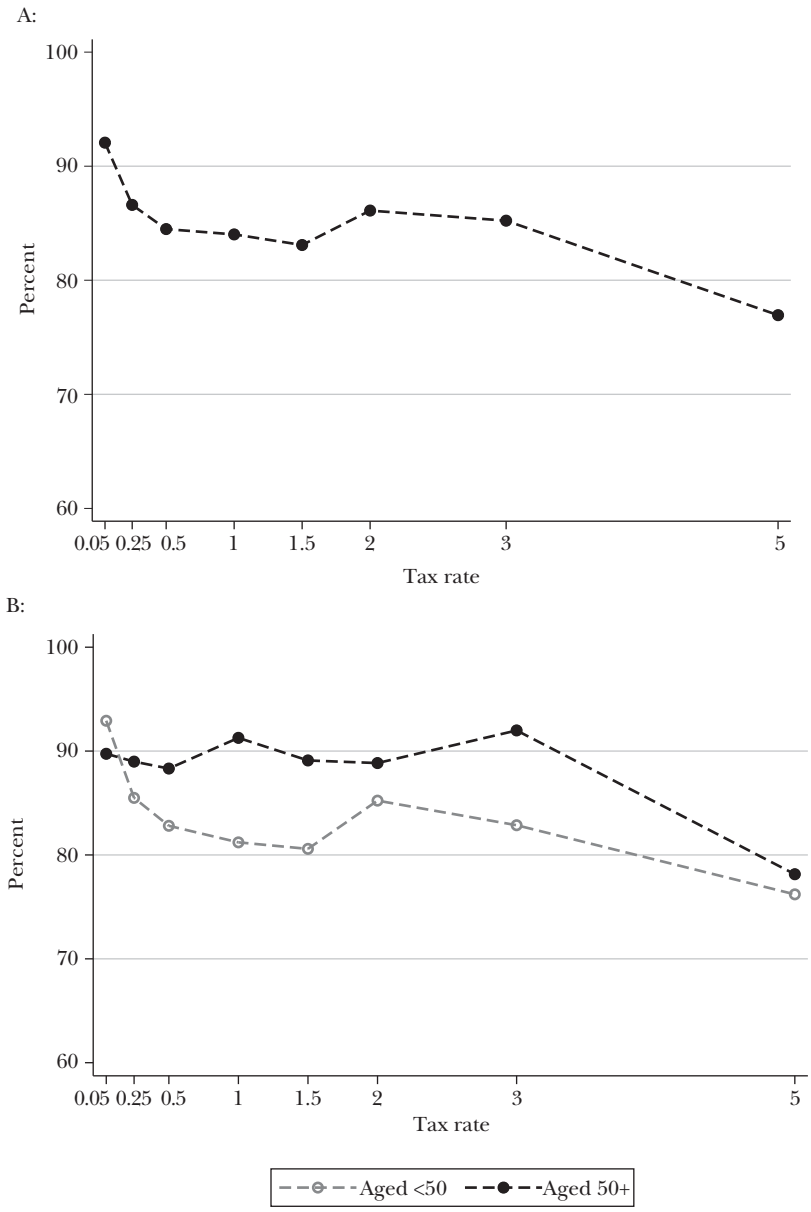
One strategy to reduce moral hazard is to increase insurance premia. But if the system is compulsory for self-employed workers, then, as argued above, this may crowd out self-employment in relatively low paid services. If instead the system is voluntary, then raising contribution rates for the self-employed may create a problem of adverse selection, whereby only workers with higher risk of unemployment subscribe. The experience of Sweden after a hike in contribution rates suggests that it was mainly those facing a lower risk of long-term unemployment who left the scheme (Kolsrud 2018).

A hypothetical valuation experiment that we carried out in the Italian surveys confirms adverse selection may be an important issue. A hypothetical discrete choice experiment was set up in the survey through offering respondents a "vignette" style choice of different scenarios regarding sick pay so as to elicit willingness to pay.[11] Respondents were asked to choose between two otherwise identical jobs: the first with no paid sick leave coverage, the second with paid sick leave provided by social security conditional on social insurance contributions of a given percentage of their gross monthly income, ranging across eight randomly chosen values from 0.05 to 5 percent. Such percentage was varied randomly across individuals. By plotting the percentage of respondents choosing the contract with paid sick leave coverage at any given level of the contribution rate, we can trace a willingness to pay or demand curve for paid sick leave. Results are reported in Figure 2. As we would expect, the curve is downward sloping and points to relatively high levels of willingness to pay for paid sick leave, with approximately 85 percent of respondents willing to pay a contribution rate of 0.72 percent—which was the prevailing contribution rate in Italy at the time the survey was deployed.[12] It also appears that the demand curve for the self-employed above the age of 50 is systematically above the demand curve for workers less than 50 years of age, which is suggestive of adverse selection,

[11] Vignette-based questions have been used widely in some areas of economics to assess willingness to pay for amenities (most notably in environmental economics), but rarely to date in labor economics. Some exceptions include the already discussed Mas and Pallais (2017), an internet survey in Denmark assessing willingness to pay for fringe benefits (Eriksson and Kristensen 2014), and an internet survey in India assessing willingness to pay for a job guarantee (Dhingra and Machin 2019).

[12] Due to sample size issues, tracing a separate demand curve for self-employed with and without employees leads to noisy results for the former group. From a visual inspection of the results, the two groups do not appear to have substantially different levels of willingness to pay. Detailed results of this vignette experiment are available upon request.

Figure 2
Willingness to Pay for Paid Sick Leave



Source: fRDB Survey.
Note: The graphs report the results of a hypothetical valuation experiment carried out in the Italian survey. Respondents were asked to choose between two otherwise identical jobs, the first with no paid sick leave coverage and the second with paid sick leave coverage conditional on social insurance contributions of a given percentage of their gross monthly income. Such percentage was varied randomly across individuals. The randomized contribution rate could take the following values: 0.05, 0.25, 0.5, 1, 1.5, 2, 3, or 5 percent. The graphs plot the percentage of respondents choosing the contract with paid sick leave coverage at any given level of the contribution rate, that is, the empirical demand curve for paid sick leave. Panel A reports results pooling all self-employed workers. Panel B reports results separately for individuals aged less than 50 (black circles) and aged 50 and over (hollow circles).

although income effects (likely to be higher for older workers) may also contribute to explain this result.

Final Remarks

Solo self-employment accounts for between 4 and 22 percent of total employment in the countries of the OECD area. It has been rising relative to self-employment with dependent workers in most countries and rising in absolute terms in almost half of the countries. However, we still know little about the nature of these jobs, the way they interact with wage setting, or the welfare gains and losses associated with their development. This paper begins the task of filling this gap by drawing on ad hoc surveys carried out in the United Kingdom, the United States and Italy, and on secondary individual-level and country-level data sources.

Although these three countries have quite different labor market institutions, historical levels of self-employment, and recent unemployment dynamics, some of the patterns we find are remarkably similar. Solo self-employment appears to be an intermediate category between employment and unemployment. It shares important characteristics with underemployment. In particular, many workers are hourly and liquidity constrained and earn less than workers in traditional jobs and in self-employment with employees, even on an hourly basis. Moreover, a substantial share of solo self-employed workers are vulnerable to idiosyncratic shocks because a single client provides more than 50 percent of their earnings.

The income insecurity that these workers face, together with the fact that they typically have few (if any) employment rights, creates a strong demand for social protection. However, designing such a program raises hard questions. Introducing social insurance programs where the self-employed make contributions on a voluntary basis would pose problems of adverse selection. Making the contributions compulsory and costly in order to reduce moral hazard may drive some of the self-employed—and in some cases their employees as well—out of work. It would also increase the liquidity constraints of the self-employed remaining in business.

In designing employment and tax policies, policymakers should reduce the incentives to hide what are de facto dependent employment positions under self-employment conditions. One example of distorted incentives is the case of employers who tilt the contractual composition of their workforce towards nondependent employment in order to avoid minimum wage and employment protection legislation. Another example is the more favorable tax treatment that many countries have of self-employment vis-à-vis dependent employment and that distorts individual incentives to sort into self-employment and firms' incentives to hire under traditional employment contracts. In this respect, reforms in the direction of preventing minimum wage or employment protection legislation avoidance and equalizing differential tax treatment ought to be considered. Finally, even relatively light exclusivity clauses—preventing the worker from supplying labor to other employers—should be carefully monitored and possibly banned if they strengthen

the monopsony power of the firm in using the services of gig workers and limit the use of self-employment as an income smoothing device by workers. Similar considerations apply to "no compete" or "no poaching of workers" agreements which are becoming increasingly pervasive in the US labor market (Krueger and Ashenfelter 2018). Whilst predominantly applied to employees, such clauses appear to be extended to freelance workers, too.

Our findings and conclusions should be further tested over a larger variety of settings and institutional configurations. One possibility is by running similar surveys in other countries and through time. Another research area is the development of methods for measuring the extent of labor market slack, particularly in light of the observation that the conventionally used unemployment rate has become increasingly narrow in its inability to pick up various aspects of underemployment that have acted to dampen wage growth in the recent past. Our surveys suggest that measures of labor market slack could usefully be refined to take into account the hours-constrained features of some of the new solo self-employment and other types of alternative work arrangements that have become increasingly prominent in contemporary labor markets.

■ *We are grateful to Saverio Bombelli and Paolo Naticchioni for assistance with the design of the Italian survey, to Nikhil Datta for useful feedback on the design of the UK survey, and to Kevin Deluca for help with the analysis of the US survey data. Henriette Druba and Ivan Lagrosa provided excellent research assistance. We acknowledge financial support from Fondazione Rodolfo Debenedetti, from the LSE Centre for Economic Performance's "Informing the Industrial Strategy" project (ESRC ES/S000097/1), and from the Turing-HSBC-ONS Economic Data Science Award.*

This paper is dedicated to the memory of Alan Krueger, our friend and colleague, who passed away on March 16, 2019.

References

Abraham, Katharine G., and Ashley Amaya. 2019. "Probing for Informal Work Activity." *Journal of Official Statistics* 35: 487–508.

Abraham, Katharine G., John C. Haltiwanger, Kristin Sandusky, and James R. Spletzer. 2018. "Driving the Gig Economy." https://www.irs.gov/pub/irs-soi/18conpansandusky.pdf (accessed December 8, 2019).

Abraham, Katharine, John Haltiwanger, Kristin Sandusky, and James Spletzer. Forthcoming. "Measuring the Gig Economy: Current Knowledge and Open Issues." In *Measuring and Accounting for Innovation*

in the 21st Century. Chicago: University of Chicago Press.
Audretsch, David B., Max C. Keilbach, and Erik E. Lehmann. 2006. *Entrepreneurship and Economic Growth*. New York: Oxford University Press.
Bell, David N.F., and David G. Blanchflower. Forthcoming. "Underemployment in the US and Europe." *Industrial and Labor Relations Review*.
Blanchflower, David G., and Andrew J. Oswald. 1995a. "An Introduction to the Wage Curve." *Journal of Economic Perspectives* 9 (3): 153–67.
Blanchflower, David, and Andrew Oswald. 1995b. *The Wage Curve*. Cambridge: MIT Press.
Boeri, Tito. 2018. *Fondazione Rodolfo Debenedetti Survey of Independent Workers*. Milano: fRDB.
Bureau of Labor Statistics and Census Bureau. 2019. *Current Population Survey, Basic Monthly Data*. Washington, DC: BLS and Census Bureau.
Datta, Nikhil, Giulia Giupponi, and Stephen Machin. Forthcoming. "Zero Hours Contracts and Labour Market Policy." *Economic Policy*.
Dhingra, Swati, and Stephen Machin. 2019. "The Value of a Job Guarantee." Unpublished.
Eriksson, Tor, and Nicolai Kristensen. 2014. "Wages or Fringes? Some Evidence on Trade-Offs and Sorting." *Journal of Labor Economics* 32 (4): 899–928.
Evans, David S., and Boyan Jovanovic. 1989. "An Estimated Model of Entrepreneurial Choice under Liquidity Constraints." *Journal of Political Economy* 97 (4): 808–27.
Farrell, Diana, Fiona Greig, and Amar Hamoudi. 2018. *The Online Platform Economy in 2018: Drivers, Workers, Sellers, and Lessors*. Washington, DC: JP Morgan Chase Institute.
Farrell, Diana, Fiona Greig, and Amar Hamoudi. 2019. "The Evolution of the Online Platform Economy: Evidence from Five Years of Banking Data." *AEA Papers and Proceedings* 109: 362–66.
Giupponi, Giulia, and Stephen Machin. 2019. *Survey of Self-employment and Alternative Work Arrangements 2018*. Colchester: UK Data Service.
Hanauer, Nick, and David Rolf. 2015. "Shared Security, Shared Growth." *Democracy* 37 (Summer).
Harris, Seth D., and Alan B. Krueger. 2015. "A Proposal for Modernizing Labor Laws for Twenty-First-Century Work: The 'Independent Worker.'" Hamilton Project Discussion Paper 2015-10.
Heckman, James J. 1981. "Heterogeneity and State Dependence." In *Studies in Labor Markets,* edited by Sherwin Rosen, 91–140. Chicago: Chicago University Press.
Henley, Andrew. 2004. "Self-Employment Status: The Role of State Dependence and Initial Circumstances." *Small Business Economics* 22 (1): 67–82.
Hong, Gee Hee, Zsóka Kóczán, Weicheng Lian, and Malhar Nabar. 2018. "More Slack Than Meets the Eye? Recent Wage Dynamics in Advanced Economies." IMF Working Paper 18/50.
Hurst, Erik, and Benjamin Wild Pugsley. 2011. "What Do Small Businesses Do?" *Brookings Papers on Economic Activity* 2 (Fall): 73–142.
Hyslop, Dean R. 1999. "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women." *Econometrica* 67 (6): 1255–94.
Istituto Nazionale di Previdenza Sociale (INPS). 2018. *XVII Rapporto Annuale*. Rome: INPS.
Istituto Nazionale di Statistica. 2019. *Rilevazione Continua delle Forze di Lavoro*. Rome: ISTAT.
Jackson, Emilie, Adam Looney, and Shanthi Ramnath. 2017. "The Rise of Alternative Work Arrangements: Evidence and Implications for Tax Filing and Benefit Coverage." Office of Tax Analysis Working Paper 114.
Jovanovic, Boyan. 1994. "Firm Formation with Heterogeneous Management and Labor Skills." *Small Business Economics* 6 (3): 185–91.
Katz, Lawrence F., and Alan B. Krueger. 2017. "The Role of Unemployment in the Rise in Alternative Work Arrangements." *American Economic Review: Papers and Proceedings* 107 (5): 388–92.
Katz, Lawrence F., and Alan B. Krueger. 2018. "The Rise and Nature of Alternative Work Arrangements in the United States, 1995–2015." *Industrial and Labor Relations Review* 72 (2): 382–416.
Katz, Lawrence F., and Alan B. Krueger. 2019. "Understanding Trends in Alternative Work Arrangements in the United States." *Russell Sage Foundation Journal of the Social Sciences* 5 (5): 132–46.
Kolsrud, Jonas. 2018. "Sweden: Voluntary Unemployment Insurance." In *The Future of Social Protection: What Works for Non-standard Workers?*, 197–224. Paris: OECD Publishing.
Koustas, Dmitri K. 2018. "Consumption Insurance and Multiple Jobs: Evidence from Rideshare Drivers." https://uchicago.app.box.com/v/DKoustas-RideSmoothing-WP (accessed December 8, 2019).
Krueger, Alan B. 2017. *Princeton Self-employment Survey*. Princeton: Princeton University.
Krueger, Alan B. 2018. "Independent Workers: What Role for Public Policy?" *Annals of the American Academy of Political and Social Science* 675 (1): 8–25.

Krueger, Alan B., and Orley Ashenfelter. 2018. "Theory and Evidence on Employer Collusion in the Franchise Sector." NBER Working Paper 24831.

Lazear, Edward P. 2004. "Balanced Skills and Entrepreneurship." *American Economic Review* 94 (2): 208–11.

Levine, Ross, and Yona Rubinstein. 2017. "Smart and Illicit: Who Becomes an Entrepreneur and Do They Earn More?" *Quarterly Journal of Economics* 132 (2): 963–1018.

Mas, Alexandre, and Amanda Pallais. 2017. "Valuing Alternative Work Arrangements." *American Economic Review* 107 (12): 3722–59.

Northern Ireland Statistics and Research Agency. 2019. *Quarterly Labour Force Survey*. Newport: Office for National Statistics

OECD. 2015. *In It Together: Why Less Inequality Benefits All*. Paris: OECD Publishing.

OECD. 2018. *Job Creation and Local Economic Development 2018: Preparing for the Future of Work*. Paris: OECD Publishing.

OECD. 2019. *OECD Employment Outlook 2019: The Future of Work*. Paris: OECD Publishing.

Parker, Simon C. 2004. *The Economics of Self-Employment and Entrepreneurship*. Cambridge: Cambridge University Press.

Spreitzer, Gretchen M., Lindsey Cameron, and Lyndon Garrett. 2017. "Alternative Work Arrangements: Two Images of the New World of Work." *Annual Review of Organizational Psychology and Organizational Behavior* 4 (1): 473–99.

Taylor, Matthew. 2017. *Good Work: The Taylor Review of Modern Working Practices*. London: Department for Business, Energy and Industrial Strategy.

Tobsch, Verena, and Wemer Eichhorst. 2018. "Germany: Social Insurance for Artists and Writers." In *The Future of Social Protection: What Works for Non-standard Workers?*, 123–43. Paris: OECD Publishing.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 196–221

The Economics of Maps

Abhishek Nagaraj and Scott Stern

For centuries, human beings have codified their geographic knowledge in maps. Mapmaking was a large and economically significant activity during the Middle Ages, and new maps were a central tool leveraged by explorations undertaken during the Age of Discovery. More recently, digital maps of the world, such as Google Maps, have been among the important applications of digital technology. Digital maps have not only enabled access to real-time transportation and traffic information, but have also supported location-based innovations such as ridesharing apps, real estate portals, and local search engines—and are a core input into the $340–400 billion dollar geospatial technology and location intelligence industry.

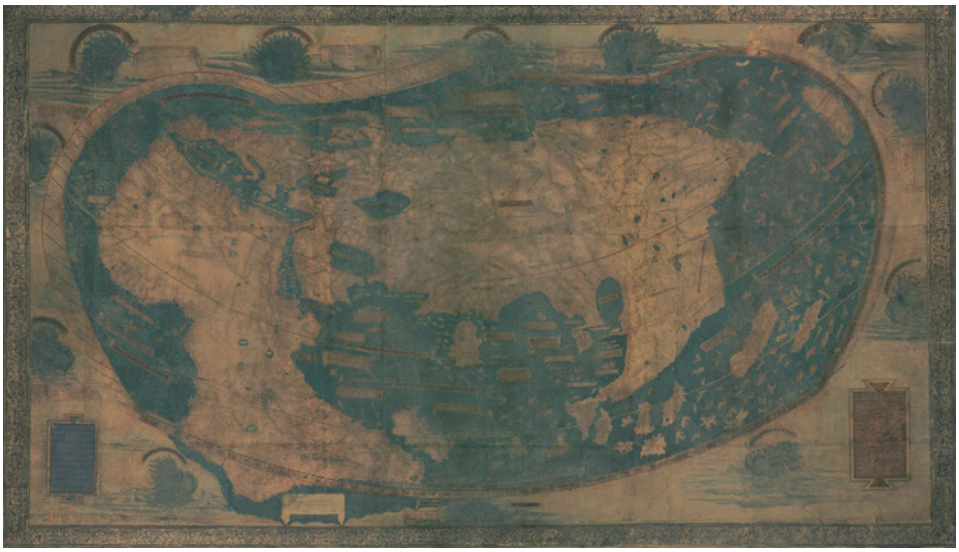
Consider how mapmakers influenced the choices and explanations of explorers via the example of one of the most famous maps ever produced. The *Martellus Map* was a Mappa Mundi (a medieval European world map) by Henrich Martellus, a geographer and cartographer from Nuremberg, who lived and worked in Florence from 1480 to 1496. While the Martellus Map was relatively accurate, it deviated to some extent from other maps of its day. The southern tip of Africa was extended to 45 degrees south latitude (even though it is actually at 34 degrees). It also extended the entire east-west length of the Eurasian landmass (from 180 degrees to 240 degrees). These miscalculations supported a theory that Cipangu

■ *Abhishek Nagaraj is Assistant Professor in the Management of Organizations, University of California at Berkeley, Haas School of Business, Berkeley, California. Scott Stern is David Sarnoff Professor of Management, MIT Sloan School of Business, Cambridge, Massachusetts and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are nagaraj@berkeley.edu and sstern@mit.edu.*

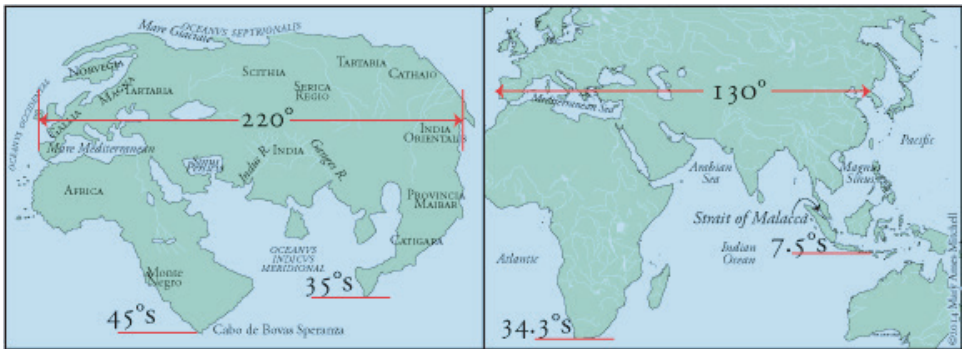
For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.196.

Figure 1
Martellus Map of the World (circa 1489) and Its Distortions

A: The Martellus Map



B: Stylized representation



Source: Panel A: Henricus Martellus. "Maretllus' World Maps." 1489–1490. Last updated August 15, 2019. http://www.myoldmaps.com/late-medieval-maps-1300/256-henricus-martellus/. Panel B: Mary Ames Mitchell. "Columbus' New Proposal." Last updated 2015. http://www.crossingtheoceansea.com/OceanSeaPages/OS-62-ColumbusNewProposal.html.

(Japan) was significantly closer to the west of Europe than it actually is (Davies 1977). While "ground truth" would indicate that a route going to Japan via Africa was considerably shorter, the Martellus Map made a westward voyage to Japan seem attractive, as illustrated in Figure 1. The Martellus Map described a view of the world that may have shaped the course of history through the unanticipated discoveries of the North American continent by European explorers. Critically, it is believed to have been referenced by Christopher Columbus in planning his voyages, was used to support the financing of his expedition, and was ultimately the basis for his

mistaken belief that he had discovered India when in fact he was in the Bahamas (Vietor 1963).

In modern empirical work in economics, maps play an important role as data sources (Glaeser et al. 1992; Moretti 2012; Naik, Raskar, and Hidalgo 2016; Chetty et al. 2014; Dell 2010), but economists have rarely undertaken the systematic study of the production of maps as a knowledge good and their consequences for economic and social outcomes. However, a recent flurry of work across disparate subfields has begun to remedy this gap and includes work that looks at the impact of satellite mapping (Casaburi and Troiano 2016; in this journal, Donaldson and Storeygard 2016; Katona et al. 2018; Nagaraj 2018), local business maps (Luca, Nagaraj, and Subramani 2019), subway maps (Larcom, Rauch, and Willems 2017), redlining maps (Aaronson, Hartley, and Mazumder 2017), and flood insurance maps (Michel-Kerjan 2010). In addition to these systematic studies, maps potentially play a role in urban economics; industrial organization through locations of firms and customers; public finance via topographical, census, tax, insurance, and weather maps; political economy (via policies on gerrymandering and property rights); and housing and financial markets. The connections between maps and these topics remain largely uncharted territory.

The present essay seeks to provide a theoretical lens to unify recent work on the role that geographic information plays in economic geography. We review and unify a variety of studies in different literatures that serve to establish the causal role that maps play in shaping economic outcomes. As context, we also provide a brief overview of the multi-billion dollar mapping and geospatial sector of the economy. Building on insights from cartography, we then argue that maps are composed of data and designs, serving as a novel type of information good with unique and specific properties. We then outline the economic properties of maps in terms of fixed costs, rivalry, and excludability and trace out implications for the social versus private returns to mapmaking. This exercise helps clarify possible market failures in mapping supply and the role of the public sector in this area.

We then explore the economic implications of a central insight from cartography that "a map is not the territory" (Korzybski 1933, 750). Maps are fundamentally a representation of physical space different from ground truth. We argue that representations appearing on a map are not an objectively "best" way to represent a geography, but instead reflect the goals, incentives, constraints, and choices of map producers, which themselves depend on particular economic and strategic environments. We endogenize the process of cartographic representation and clarify key economic dimensions which influence representational choices. In particular, we examine: (1) the costs of mapmaking, (2) the nature of demand for maps, (3) intellectual property and the competitive environment, (4) the role of innovation in mapmaking technology, and (5) incentives of mapmaking organizations or individuals. We offer predictions about how these factors shape the ways in which maps may differ from ground truth, and the economic and social consequences of these choices. We also clarify that mapmaking is a dynamic, endogenous process subject to path dependence, and that these five dimensions provide sources of exogenous variation to this path-dependent trajectory. We

conclude with an overview of the open theoretical and empirical research questions in this area.

The Economic and Social Consequences of Maps

Why study maps? Even before their present-day relevance, maps have played an important—albeit unintended—role in shaping history. These changes have occurred not only because maps provide useful information, but also because they distort and represent such information in consequential ways. During the US Civil War, General George McClellan's reliance on a distorted map, one which failed to show the Warwick River as a significant obstacle to an invasion of Richmond, resulted in the war being unnecessarily prolonged, producing a hefty loss of lives on both sides (Shulten 2012; Monmonier and de Blij 1996). Other consequences of inaccurate maps were more deliberate. British colonialists justified their ownership of some territories by employing colors and symbols that represented regions of India as British possessions, even though their control on the ground was tenuous and far from complete. Such maps helped to encourage further investments by the British government in securing India for the British Empire, providing significant rewards for the colonialists (Barrow 2004).

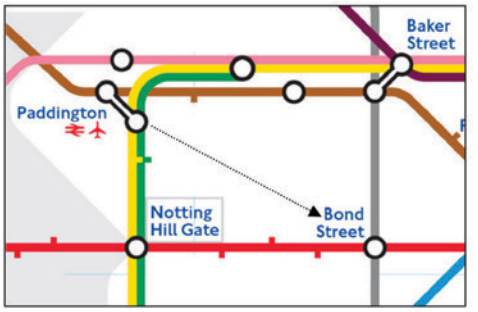
Emerging empirical literature in economics and related fields also points towards the causal role that mapping and geographic information has played in shaping social and economic outcomes. Given the endogenous nature of maps and the variety of factors that systematically shape them, economists have tended to exploit shocks to the quality of maps (coming from innovations in mapmaking or spatial variation in their accuracy) to identify empirically their causal role. We provide a brief overview of these studies in a variety of different fields.

Consider the case of the 2014 London Underground strike. Service stoppages prompted regular riders to consider alternative commuting routes, at which point they discovered that their previous choices had been suboptimal (Larcom, Rauch, and Willems 2017). While they primarily focus on how agents learn about optimal routing, a key finding shows a larger proportion of passengers found they were engaged in suboptimal routing in areas where the Tube map was more distorted. For example, as shown in Figure 2, a traveler going from Paddington to Bond Street stations had a choice of transferring via either Baker Street or Notting Hill Gate. Though the Notting Hill Gate route was in fact 15 percent slower on average, more than 30 percent of passengers used this route simply because the London Tube map displayed the Notting Hill Gate station as south rather than west of Paddington, causing the total length of the two routes to falsely appear equal (Guo 2011). It was only because real-world experimentation was induced that commuters learned about the mistaken inferences from the canonical yet inaccurate London Tube map.

Beyond effects on individual decision-making, mapmaking distortions affect a broad range of areas, including public finance. Property taxes have traditionally depended on the codification of property rights through parcel maps. Incomplete

Figure 2
An Example of How the London Tube Map Distorts Distances

A: Schematic Tube map



B: Geographical map

Source: Adapted from Guo (2011). Panel A: London Underground. Panel B: Simon Clarke.

or inaccurate parcel maps may result in misspecification of property lines, offering the potential for tax avoidance. Casaburi and Troiano (2016) assess how the use of satellite imagery in Italy allowed for improved parcel maps, resulting in the identification of more than 2 million "ghost buildings," facilitating a crackdown on tax evasion and ultimately enhancing tax revenues by €472 million over a four-year period. Similarly, during the Greek debt crisis, the Greek government leveraged satellite imagery from Google Maps to detect undeclared property improvements such as swimming pools. In the suburbs of Athens alone, the government's count of swimming pools rose from 324 to 16,974 after maps were deployed for this use (Daley 2010).

Similarly, the pricing and demand for flood insurance both depend on the information presented in flood risk maps (Michel-Kerjan 2010). The National Flood Insurance Program's (NFIP) reliance on outdated maps resulted in Colorado policyholders paying 15 times more in premiums compared to claims, while Mississippi policyholders received five times more in claims than they paid in premiums. Inaccurate flood maps also affect the choices homeowners make about their levels of insurance coverage. Families in New Orleans underestimated their levels of flood risk and therefore underinvested in insurance protection, a choice that proved costly in the wake of Hurricane Katrina.

Other recent work highlights the potential influence of maps on investments in regional natural resources, which in many analyses are assumed to be exogenous and known. But as highlighted by Wright (1990), US leadership in energy and mineral resources is not simply a product of natural endowments, but also relies on systematic investments in the topographic and geological mapping of regions through organizations like the US Geological Survey. The effect of investments in mapping is clearly demonstrated by the history of gold exploration and discovery. The introduction of satellite imagery by the NASA Landsat program during the 1970s facilitated the identification of geographical lineaments that strongly predict

the presence of gold deposits. Nagaraj (2018) takes advantage of random variations in the timing and quality of these images (like whether the image was cloud-free) to demonstrate that new maps resulted in nearly doubling the likelihood of discovery of new gold deposits when compared to unmapped regions, an effect disproportionately associated with finds from smaller and younger exploration firms.

In some cases, maps can illuminate the spatial distribution of economic activity, such as research using data on nighttime lights from the US Air Force Defense Meteorological Satellite Program's Operational Linescan System (DMSP OLS) (Croft 1978; Henderson, Storeygard, and Weil 2012; Donaldson and Storeygard 2016; Baragwanath et al. 2019). In other cases, maps can affect economic outcomes in the region they aim to describe. The Home Owners' Loan Corporation, founded during the Great Depression to regulate the housing market, created residential security maps to assess the risk of lending in a specific location. Districts deemed to have lower residential security were "redlined," resulting in higher racial segregation and lower homeownership rates, credit scores, and house values in subsequent decades (Aaronson, Hartley, and Mazumder 2017). These maps not only reflected the existing reality of segregation but served as a tool for the state to exacerbate discriminatory practices (Scott 1999).

Finally, there are a number of consequences when maps establish political boundaries. Many of these consequences are unintended, especially because the mapmakers themselves have different reasons for the ways in which they define borders or label areas. For example, when the 49th parallel was chosen as the dividing line between the United States and Canada and ratified in the Oregon Treaty in 1846, the vague language regarding the channel around Vancouver Island led to an armed standoff between 1859 and 1872 when arbitration awarded the San Juan Islands to the United States (Kershner 2013). Some international disputes are not so easily resolved. The Sykes-Picot Agreement drafted during World War I divided the Ottoman Empire into new states using a ruler, resulting in imprecise and arbitrary boundaries that have arguably been at the heart of the instability of that region for the last century (Wright 2016). Maps seem to have an outsized role in shaping outcomes of interest to economists and other social scientists.

The Geospatial Industry

A first step to uncovering the economics of maps is to understand the industrial organization of the geospatial industry consisting of organizations that gather, store, process, analyze, and distribute geospatial information. Consumer-facing mapping services include digital mapping technologies such as Google Maps, which by itself has over 1 billion active monthly users globally and over 150 million active users in the United States (Popper 2017; Clement 2018). Digital mapping services are highly valued by these consumers; for example, Brynjolfsson, Collis, and Eggers (2019) use choice experiments to estimate that the median US consumer would have to be paid at least $3,648 in order to forego digital maps for a year (exceeding the value associated with digital video, social media, or messaging). In addition to end users, the mapping industry serves diverse organizations and stakeholders across a wide variety of industries (mining, agriculture, insurance, and real estate, to name a few)

and public sector organizations (from local to national governments). The broad reach of the geospatial industry is associated with significant economic output. The size of the global geospatial industry is estimated to be between $339–400 billion (Geospatial Media and Communications 2019; AlphaBeta 2017), with a somewhat older estimate just for the United States of about $75 billion (Boston Consulting Group 2012). Even if one focuses more narrowly on the "surveying and mapping" sector (NAICS code 54137), geospatial data collection involves 16,800 businesses with total revenue of $7.8 billion in 2018 in the United States (O'Connor 2018).

Figure 3 divides the industry into four broad groups: geospatial technology providers (hardware, earth observation, software), data providers (surveying and mapping companies, government), delivery platforms (business-to-business and business-to-consumer) and analytics (business-to-business, consulting and design agencies). While some sectors such as geospatial technology are relatively fragmented (for example, there is no single dominant surveying company), several key areas are highly oligopolistic (location-based mapping services) and others feature a single dominant firm (such as ESRI in the area of geographic information system software). There are also a wide range of business strategies across and within these sectors. While some firms such as Rand McNally historically sold maps of their own design directly to consumers, other companies such as Mapbox license mapping data and software to customers to build maps of their own design (and for their own purposes). Also, some leading companies such as Google employ a platform strategy in which maps are provided for free to consumers whose use is then monetized through location-sensitive and context-sensitive advertising. Google Maps, for example, is expected to generate revenue to the tune of $11 billion by 2023 primarily through advertising (Schaal 2019). The demand for mapping products seems to be growing rapidly due to the growth in automation, artificial intelligence, and advanced analytics increasing the adoption of geospatial technologies in organizations (BCG 2012; Geospatial Media and Communications 2019).

The Production of Maps

We now turn to describing the essential elements of mapmaking, as a first step to uncovering the economics of mapping information. At its core, a map takes selected attributes attached to a specific positional indicator (spatial data) and pairs it with a graphical illustration or visualization (design) (DiBiase 2008). The canonical political "world map" visualizes spatial data about country names and political boundaries, while a tourist map might visualize data on the location of historical monuments along with walking trails and bus routes. While the scope of mapmaking is quite broad (ranging from weather forecasts to the identification of historical battlefield locations), mapmaking is but a subset of the broader realm of knowledge production (for example, it excludes scientific discoveries such as electromagnetism as well as creative work such as novels). Maps are meaningful because they are associated with a specific terrain, but they are not intended to provide a full or comprehensive description of the underlying reality. Instead, it is well understood that even the most "complete" maps are only abstractions

Figure 3
An Overview of the Geospatial Industry

Category	*Type*	*Leading organizations*	*Selected size estimates*	*Competitive*	*Public/ private*
Geospatial technology providers	Hardware	Airbus, Boeing, Lockheed Martin, Raytheon	€42B estimated worldwide revenues in 2015 (European GNSS Agency 2017).	Concentrated with diverse periphery	Mixed
	Remote sensing satellites	Maxar/ DigitalGlobe, Planet Labs, governments	MDA purchased satellite imaging company DigitalGlobe for $2.4B in 2017 (MDA Corporation 2017).	Concentrated	Mixed
	Software	Esri, Pitney Bowes, QGIS	Esri's revenue was $1.1B in 2014 (Helft 2016).	Concentrated	Private and open source
Data providers	Organizations	NAVTEQ/ HERE, TomTom/ TeleAtlas, OpenStreetMap, USGS/NASA, UK Ordnance Survey	Location-based data generated an estimated $230B in worldwide revenue in 2016 (AlphaBeta 2017). In FY 2019, the US government spent $1.4B on defense GPS and $96M on civil GPS augmentation (GPS.gov 2019).	Concentrated	Mixed
	Surveying and mapping companies	No major national/ international players	$7.8B from 16,800 firms in the US in 2018 (O'Connor 2018).	Competitive	Private
Delivery platforms (B2C/B2B)	Apps/ location-based services	Google Maps, OpenStreetMap, Mapbox	Google purchased social navigation app Waze for $1.1B in 2013 (Lunden 2013).	Concentrated but diversifying	Private
Analytics (B2B)	Consulting and design agencies	BCG's GeoAnalytics group, terraPulse, Farmers Edge	Global market size estimated at $78.6B in 2019 (Geospatial Media and Communications 2019).	Competitive	Private

Source: Authors.

or incomplete descriptions of the underlying reality (Robinson et al. 1995; Monmonier and de Blij 1996).

Maps are not made at random but by mapmakers who exercise significant discretion and agency, whose choices are shaped by the economic, strategic, and institutional environment in which a particular map is produced. Two key elements of mapmaking are worthwhile to distinguish: the gathering and organizing geospatial information (data) and, conditional on that data, the use of geospatial tools and visualizations to create a particular map (design).

The first step in any cartographic production is finding a data source that includes both the geographic locations of interest and the associated attributes of interest to a mapmaker (DiBiase 2008). For example, a cartographer interested in making a map of restaurants near a tourist attraction must first acquire the latitude and longitude locations of the hotels of interest, associated attributes (for example, three-, four-, or five-star status) as well as some information for the "base map," which refers to the location of key highways, towns, political boundaries, and other key background. Base-mapping data can come from different places, including free and public sources (like the US Geological Survey) as well as private sources (such as Google Maps). Data on the object of interest can sometimes be obtained through an open-source or public initiative, but might need to be licensed or even directly collected at significant cost. The eventual map and its informativeness is inherently constrained by the choice of data provider. For example, Yelp maps rely on external data aggregators for data on local business listings and such providers often miss listings for businesses that are in more remote locations or smaller in size. In fact, when compared with administrative data from tax records, Yelp coverage is found to be in the range of about 60 percent (Luca, Nagaraj, and Subramani 2019), although such gaps in coverage can improve almost overnight when data providers add missing listings to their database. Incomplete or selective data can be consequential; in the case of Yelp listings, the exclusion of a restaurant is estimated to reduce restaurant revenue to the tune of 5–12 percent.

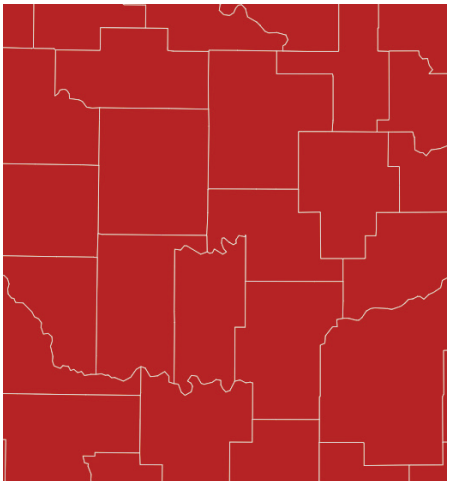
Having chosen data sources and selected key locations and attributes, the mapmaker must then pick a design that visualizes the underlying information. This process is based around a wide variety of choices, including those around simplifying certain features and exaggerating others, using symbols and classifying attributes into groups, and so on. A prominent example of a design choice is aggregation, which involves deciding the geographical unit at which information is displayed. Consider alternate maps for the 2016 presidential election in a predominantly Republican area, such as the areas of Oklahoma shown in Figure 4. A cartographer might group electoral results by county, depicting a state where all regions appear to be staunchly Republican, or the cartographer might group them by precinct, which might reveal certain pockets (like parts of Oklahoma City, Langston, or Boley) that voted for the Democratic candidate. Similarly, another consequential design choice is around mathematical projections used to represent a three-dimensional earth on a two-dimensional surface. The standard choice to adopt the Mercator projection (invented in the 16th century to aid navigation) increases the relative size of areas far from the equator, thereby increasing the perceived importance of areas such as Western Europe at the expense of large land masses at the equator, most notably Africa.

Private versus Social Returns to Mapmaking

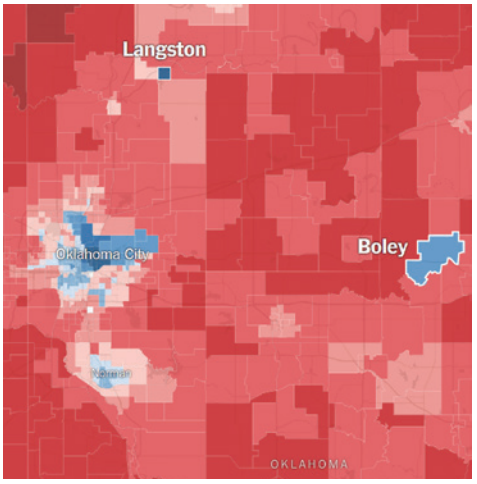
Conceptualizing maps as a design representing data has important implications for the economic properties of maps. To a first approximation, both data and designs are types of knowledge goods and so can be characterized as "non-rival" (use by one person does not preclude use by others) and partially "excludable" (it

Figure 4
Vote Patterns for 2016 US Presidential Election around Oklahoma City, OK

A: County level



B: Precinct level



Source: Panel A: Politico 2016. Panel B: Upshot Staff 2018.
Note: The darker shades of red denote majority Republican vote share, and darker shades of blue denote majority Democratic vote share.

is possible to limit use for those without explicit permission). This characterization allows us to consider the likely distortions that arise in terms of the private incentives to produce and disseminate data and designs, respectively.

First, mapping data is in many respects a classical public good. Almost by definition, mapping data is non-rival insofar as the use of data for a map by any one person does not preclude its use by others; moreover, the information underlying a given database is non-excludable because copyright law does not protect the copying of factual information. While the precise expression included within a database can be protected through copyright, the underlying geographical facts reflected in the database cannot be protected. As such, there is no means by which a data producer can preclude others from undertaking independent verification and use of a given body of geographical information (often at much lower cost than the initial sunk cost of the initial gathering and organizing of geospatial data). The combination of non-rivalry and non-excludability of mapping data makes its production prone to private underinvestment, providing a rationale for government support. Indeed, many of the most widely used maps rely on publicly funded geospatial data, including US Geological Survey topographical maps, Census demographic information, and local land-use and zoning maps. Further, even when private sector data is available in a given domain, it often relies heavily on public databases, as is the case with weather forecasting data (Lewis 2018).

Although significant bodies of mapping data are non-excludable (at which point public provision is common), there are important cases where mapping data

is in fact excludable, either through secrecy or contract. For example, the use of high-definition maps for autonomous vehicles comes with significant restrictions on the copying of the underlying data, and image maps (such as satellite or aerial maps) are themselves protected by copyright (even though the factual information contained in these images is not subject to copyright). Mapping data that allows for excludability exhibits properties more akin to a club good than a traditional public good. Specifically, the significant fixed costs of data collection combined with relatively cheap reproducibility creates entry barriers that supports natural monopolies or oligopolistic competition. It may be efficient for only a single firm to engage in data collection and for the industry to simply license these data (under agreed-upon contractual terms) from this monopoly provider. For example, DigitalGlobe is the leading provider of high-resolution, copyrighted satellite imagery, charging significant prices for access to data (whose marginal cost of reproduction is near zero) to a variety of downstream sectors, including insurance, energy, and mining. The private market for access to raw global street-mapping data is controlled by TomTom/TeleAtlas and NAVTEQ/HERE, who engage in oligopolistic competition through licensing contracts with downstream users.

Even when excludability allows for the "private provision of a public good" (Milgrom, North, and Weingast 1990), efficiency is far from guaranteed. First, in the absence of perfect price discrimination, private entities may only provide mapping data at a high price (relative to near-zero marginal cost), reducing efficient access. Beyond pricing, the private provision of mapping data may additionally be concentrated in locations with high demand (such as urban areas) to the exclusion of less concentrated regions. For example, commercial providers of satellite imagery have vastly greater amounts of data for high-density regions (such as cities) than rural areas that might be equally interesting from an environmental point of view, and even then, cities in the developed world have much greater coverage than cities in the developing world. While such prioritized data gathering might be optimal for the monopoly provider of mapping information, exclusion from mapping databases induces social distortions among downstream users and consumers.

Conditional on the production and availability of a given body of geospatial data, maps involve a second type of knowledge good through the production of a particular map design. Like data, designs are also a knowledge good in that multiple individuals can use a particular map design (and so a design is non-rival) and the degree of excludability for a given design may vary with the institutional and intellectual property environment. With that said, a striking feature of a map design is that, almost by construction, a map is created for the purpose of visual inspection, and it is much easier to copy than a database (which might be protected by secrecy or contract). One consequence of this is that there may be underinvestment in high-quality and distinct designs for a given body of geospatial data. For example, of the 200,000 top websites using a map, 180,000 utilize the now-standard visual design of Google Maps, rather than a design of their own making (BuiltWith 2019).

A potential consequence of the non-excludability of mapping data and designs is inefficient *overproduction* of mapping products that compete with each other. Once a given map is produced for a particular location and application (say, a city-level

tourist map), copycat maps can be produced at a lower sunk cost; because demand for maps of a given quality and granularity is largely fixed, free entry based on a given map involves significant business-stealing (Mankiw and Whinston 1986). In other words, conditional on the data and the design, and in the absence of excludability, there is likely to be a commons problem where there is an oversupply of relatively homogeneous map design varieties. Perhaps the most extreme version of the commons problem for maps is the case of a "treasure map," whereby a valuable object can be located through the use of a specialized map. While a single copy of such a map might lead to efficient exploration, competitive supply of such a map will result in a (socially inefficient) race to be the first to find the buried treasure!

Map data and map design, then, are similar in that they are both subject to potential underinvestment. But, whereas map data can be combined or represented in an almost limitless number of ways (that is, there is not likely to be "overuse" of mapping data), map designs may be subject to low incentives for production of a map design of a given quality, but then be subject to overproduction due to imitative copycats.

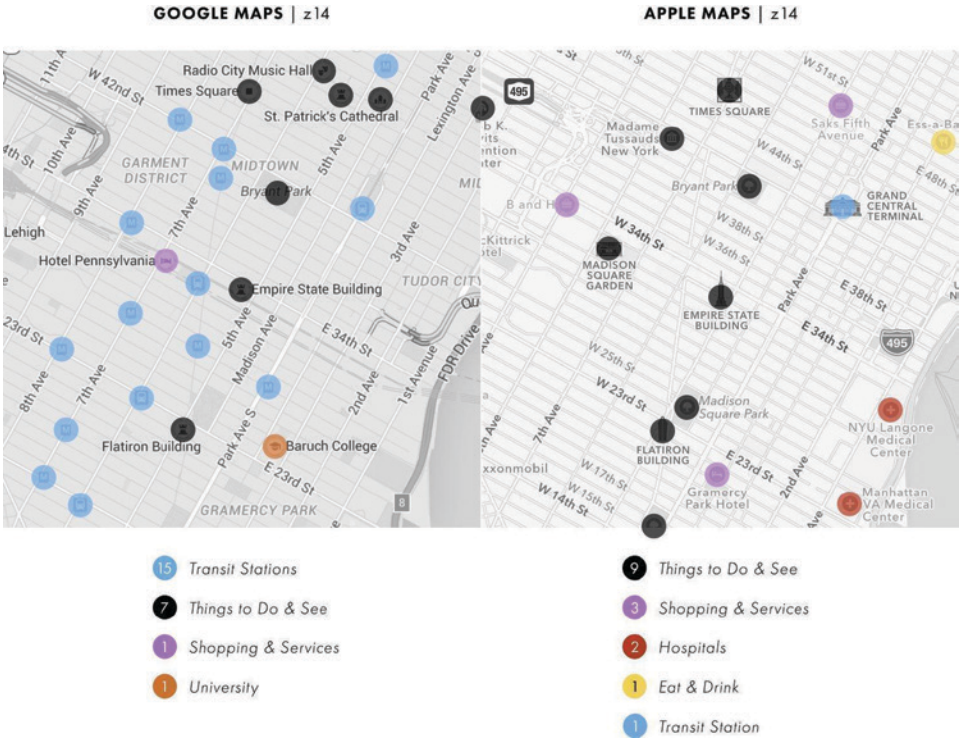
As well, though not the primary focus of our analysis, both mapping data and design choices depend on the availability of cartographic tools (from measurement instruments to design tools such as ESRI's ArcGIS software), and the availability and quality of these tools themselves depend on the institutional and intellectual property environment. Finally, it is useful to note that, beyond their functional value, maps are not only knowledge goods but also creative consumption goods, and there is an active market (and value placed) on maps with distinctive designs due to their artistry or historical significance. For example, the only known copy of the famous Waldseemuller map produced in 1507 was sold for $10 million to the Library of Congress in 2003 because this was the first map to use the name "America" and is often referred to as America's birth certificate.

How Economic and Institutional Context Affects Mapmaking

Beyond the question of possible market failures and potential remedies via the public sector in the supply of maps, the economics of mapmaking as a distinct knowledge good raises a broader question. Not only is the map *not* the territory (Korzybski 1933, 750) or a "mirror of nature" (Harley 1989), but it is also a potentially biased representation shaped by social, political, and economic forces. In other words, the choices of data and design that underly the making of a map are endogenously shaped by economic forces. The central question then becomes: how do the economic, technical, and institutional environments in which those choices are made affect the types of maps that are produced?

Consider the contrast shown in Figure 5 between the two leading mobile phone maps, Google Maps and Apple Maps. A detailed comparison of these two interactive maps for San Francisco, New York, and London by the cartographer Justin O'Beirne (2016) shows striking differences. While both Google and Apple Maps offer a similar number of features at a given level of resolution, Google Maps labels

Figure 5
A Comparison between Apple and Google Maps



Source: Adapted from O'Beirne (2016).

relatively more roadways and transit, while Apple Maps favors landmarks and shops. The differences are not small; with a given level of zoom, the average incidence of label overlap is only 10 percent. Moreover, these differences do not seem to be random. Apple's mapping priorities reflect its focus on a relatively affluent end user seeking a particular place, like the Empire State Building. Google Maps prioritizes its role as a platform for connecting map-using businesses to users, particularly through transportation applications such as Uber and Lyft. Though differences between maps are likely not noticed by most of the public, users nonetheless are presented with very different representations of an underlying territory depending on which application they use.

How might differences in the microeconomic and institutional environment affect the production of maps? We focus on the impact of variation in five critical dimensions: the costs of mapmaking, the demand for maps, competition, and intellectual property provisions, innovation, and organizational incentives. While mapmaking is a dynamic and path-dependent process, with old maps shaping newer

ones, we discuss these five dimensions as key forces that shape the quality and nature of maps in important ways.

Costs of Mapmaking

Perhaps the most important source of variation influencing mapmaking arises from dramatic variations in the costs of producing a map. Cartographic firms such as TomTom produce global maps through original surveying, and those relying on TomTom basemaps can usually use them only at significant cost. For example, in 2012, in order to launch their Maps product, Apple contracted with TomTom to license cartographic data at scale. Although Apple is only one of their clients, TomTom registers nearly $1 billion in revenue primarily from the licensing of their proprietary data (TomTom 2019). By contrast, map-based applications can also use Google Maps, usually at a lower cost but with much less flexibility in terms of selecting underlying data and choosing a custom representation. For example, while Uber has made significant investment in original mapping efforts (it planned to spend $500 million on a global mapping project as of 2016, according to Hook 2016), their consumer application largely uses a relatively generic Google Maps representation at a cost of $58 million for three years of mapping services (S-1, Uber 2019). Finally, there are also a number of open-source and relatively cheap mapmaking initiatives such as OpenStreetMap that offer a high degree of customization, but are also relatively uneven in terms of their data quality (Barrington-Leigh and Millard-Ball 2017). Tesla shifted over the 2010s towards open-source mapping technology for its in-car navigation system (Lambert 2017) and reportedly spent about $5 million for a two-year licensing deal with Mapbox in December 2015 (Bloomberg 2018). In addition to cost considerations around base maps, mapmakers also face similar choices in terms of the technology and software used to create maps as well as the cost of human capital to develop such maps. For example, an annual license per user for ESRI's ArcGIS product can cost up to $3,462 (ESRI 2016) while other tools are free.

This variation in the cost of access to mapping data as well as the cost of map design can have a significant effect of the nature of the finished map. A striking example of such cost variation comes from the Landsat program. Despite (or perhaps because of) the early success of the Landsat satellite imagery program, the US government privatized the initiative in 1984, resulting in a dramatic increase in the price of satellite data from 1984 through 1995 at which time the data was brought back into the public domain. For example, the cost to purchase one complete set of Landsat TM data covering the coterminous United States went from about $250,000 in 1982 to over $1.9 million in 1991. Nagaraj, Shears, and de Vaan (2018) explore the impacts of these cost variations on the production of scientific maps used for environmental and climate change analyses. During the high-cost privatization era, there was a much lower rate of production of high-quality environmental maps covering wide areas, particularly in the developing world. Maps based on Landsat imagery were not only less common, but they often tended to focus on narrow geographic areas rather than area-wide or country-wide surveys, in order to reduce the cost of mapping studies. However, the relative paucity of high-quality large-scale maps documenting environmental change (such as the deforestation

of the Amazon or continent-wide glacial melt) may well have delayed key scientific research and reduced the salience of some important topics in policy circles.

Cost variations can also arise from private sector mapping firms (such as Planet Labs or DigitalGlobe/Maxar in the area of satellite imagery). These firms charge significant fees and tend to serve commercial industries such as those in mining and energy, largely excluding noncommercial sectors such as academia or nonprofits who are also interested in these data. Although fees paid by the private market are not public, the US National Geospatial-Intelligence Agency signed a $44 million annual contract starting in 2019 to access Maxar (DigitalGlobe) commercial imagery (Maxar Technologies 2019). However, these firms do occasionally license their data at relatively low cost for broader social purpose. After the Haiti earthquake in January 2010, two private companies, DigitalGlobe (now Maxar) and GeoEye (acquired by DigitalGlobe in 2013), provided free, high-resolution, pre- and post-disaster satellite imagery within three days so that volunteers and experts working with the World Bank could make Building Damage Assessment maps, which are the central tools in guiding disaster relief (World Bank 2010). The availability of low-cost post-disaster imagery from private firms in the last 10 years has led to more timely and comprehensive disaster maps (which were previously based on costly aerial and on-the-ground surveys), transformed disaster mapping, and improved disaster response (Singh 2018). More generally, dramatic variation across time and space in the costs paid by mapmakers for mapping data (and complementary technology to produce maps such as software or mapping instruments) provides economists with an opportunity to trace out the supply of maps of a given type and determine the downstream consequences for economic outcomes.

Demand for Maps

Beyond cost, the nature of demand for particular types of maps will affect what maps are produced. If potential users of maps in a given territory are largely homogeneous in terms of the locations and features of interest, then different maps will likely include similar information and little differentiation in terms of design (as in the case of tourist maps). For example, the vast majority of visitors to an art museum are interested in a representation of the overall layout of the museum, key attractions (such as the Mona Lisa), and information on amenities such as bathrooms and the cafeteria. Though there are in principle an infinite number of potential representations of the Louvre, the majority of Louvre maps—including those in independently produced guidebooks—look remarkably similar given the relatively homogenous demand for information in this context.

In other contexts, demand can be quite heterogeneous across space and affect the type of maps in use. Consider the stark differences between the leading street maps of New York City versus those of Los Angeles in the pre-digital era. In New York City, the leading mapping agencies provided a relatively compact map, featuring a general overview of the territory (for example, the New York-New Jersey metropolitan region) and a small number of detailed cutouts of specific geographies like downtown Manhattan, adjacent Brooklyn locations and a separate map for midtown Manhattan listing theaters, and so on. In Los Angeles, the dominant

map was the iconic Thomas Guide, which provided a comprehensive set of detailed street-by-street maps included in an atlas-style publication weighing more than two pounds and with over 3,000 pages and was designed to be used while en route in an automobile (Daum 2015). Despite hosting populations of a similar size, heterogeneity in the nature of demand for geographic information across these two markets explains a large portion of this difference. In New York City, the most common historical use for maps was largely to navigate towards a small number of locations in Manhattan. Los Angeles, by contrast, has historically been more spread out in terms of its population and attractions, and so different users are starting from and going to a more diverse set of locations. Mapmakers responding to this variation in the heterogeneity in demand for spatial information produced a large compendium of equally detailed maps for the different regions of Los Angeles, while in New York, the standard maps had significantly greater representation and detail for certain central locations, while ignoring other regions.

Assessing the effects of heterogeneous demands on map production and use is in principle testable using methods similar to the industrial organization studies of media production and use (for example, Berry and Waldfogel 1999). A wide range of map collections have been catalogued, and the inclusion or exclusion of particular features within particular territories for a given map is potentially measurable. Finding the relationship between these differences in map production and their impacts may prove an interesting trajectory for future research.

Competition and Intellectual Property

By their nature, maps have a high fixed cost of initial development and a lower marginal cost of replication and are therefore quite sensitive to the strength of intellectual property laws protecting mapping data and representations. In fact, US copyright law from the outset offered copyright protection to "maps, charts and books," which was consistent with the idea that geographical maps were valuable forms of intellectual property that required incentives for their production and dissemination (Landes and Posner 1989). Absent perfect intellectual property rights, the production of a map encourages entry by imitative mapmakers, which reduces the incentive to produce original maps. Indeed, the explicit inclusion of charts and maps in the US Copyright Act of 1790 was motivated by the arguments of mapmakers such as Jedidiah Morse (the so-called "father of American Geography"), who argued to Congress that failure to defend his rights would result in a reduced investment in map design and production (Maher 2002).

In addition to employing copyright, firms often invest in additional strategies to protect their intellectual property. In particular, mapmakers have devised the idea of inserting fictional "paper towns" or "trap streets" in maps (Jacobs 2014). This strategy allows them to detect rivals who might copy their data (rather than collecting similar data through an original survey) and thereby protect costly investment in original data collection. Such strategies are commonly deployed by mapmakers to this day for factual data (Bridle 2012).

Our earlier discussion on non-excludability highlighted the central tension regarding the impact of intellectual property. On the one hand, an absence of

Figure 6
Competing Street-Level Imagery Maps for 639 17th Street NW, Washington, DC

A: Microsoft StreetSide

B: Google StreetView

Source: Panel A: https://binged.it/2YGYTcB. Panel B: https://goo.gl/maps/XbqqhTqSXRNY3GiW8.

formal intellectual property protection leads to underinvestment in mapping data and high-quality map design, but inefficient entry by copycat mapmakers. On the other hand, a high level of formal intellectual property protection can shift the basis of competition away from imitation and towards duplicative investment. For example, over the past two decades, no less than four different organizations—including Google Street View, Microsoft StreetSide, OpenStreetCam project, and TomTom—have undertaken comprehensive and qualitatively similar initiatives to gather street-level imagery and mapping coordinates for the entire US surface road system. While an absence of intellectual property protection might lead to underprovision, the provision of property rights for maps may instead be associated with overinvestment, as illustrated by the two very similar street-level images in Figure 6.

Finally, it is interesting to consider the nature of maps when mapmakers that do not enforce copyright, such as nonprofits or crowdsourcing communities, face competition from commercial providers who do. Nagaraj and Piezunka (2018) study crowdsourced, open maps on the OpenStreetMap platform and find that such maps are likely to look different in the presence of commercial competition as compared to cases when they are the only such platform in town. By examining how OpenStreetMap contributors respond to the entry of Google Maps in different countries around the world, they show that commercial competition causes casual mapmakers to stop contributing, while already established volunteers increase contributions. In other words, voluntary efforts to create maps may result in maps that are of high value to a small group of "superusers" but may be less aligned with overall market demand.

Innovation

Exogenous shocks from technological innovations both enable and constrain mapmakers and the mapping representations they choose. Consider the adoption of astronomical tools for navigational purposes that profoundly shaped nautical cartography in the second half of the 15th century (Ash 2007). Navigators, venturing

outside of established trade routes, incorporated tools such as the quadrant and the astrolabe (used to calculate altitudes of celestial bodies) to calculate their north-south position on the earth's surface. Before this innovation, navigators relied on portolan charts, which are maps with straight distance lines marked between points such as ports or landmarks and were designed to aid navigation by "dead reckoning" techniques (which involve navigating using distance and direction from the origin). The use of astronomical tools and mathematical navigation techniques gave birth to projected maps that use latitudes and longitudes, a system that is used to this day.

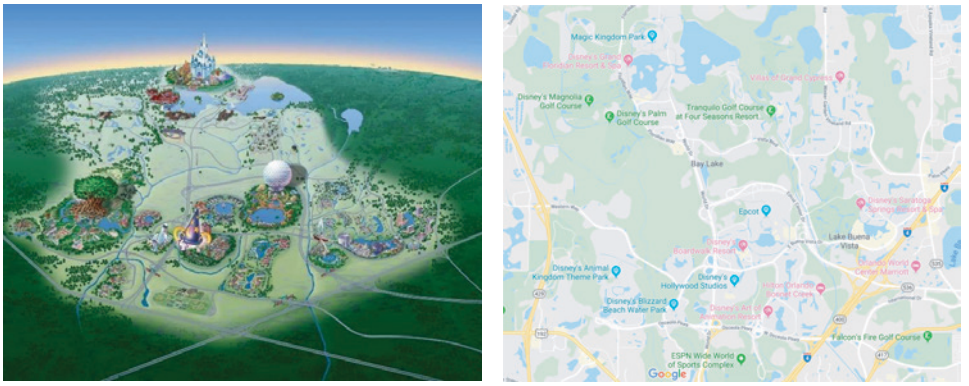
The development and adoption of new technologies continues to shape the nature of maps in the modern era. Consider the case of satellite technology discussed before. Though aerial imagery became available for significant portions of the Earth over the course of the first half of the twentieth century, systematic satellite mapping of the globe only began in 1972 with the launch of the Landsat program by the United States Geological Survey and NASA. Despite the high costs of producing these maps, the US government initially chose to distribute the underlying data (in the form of satellite photographs) at a nominal cost. Remote sensing data, such as data from satellites, allows for easier access to information and provides higher spatial resolutions and a wider geographic coverage, leading to higher quality maps in many domains, increased use by industry and, increasingly, economists (as discussed in this journal by Donaldson and Storeygard 2016).

Finally, technological shocks also provide opportunities to determine causal effects of maps. In the case of Landsat, there were significant variations in the timing of the availability of "clear" satellite maps of a given region due to differences in weather (for example, some regions were originally photographed on a cloudy rather than clear day) and luck (for example, some images were poor due to random technical errors). It was later discovered that high-quality satellite maps can be used to identify gold deposits that form at fault line locations on the surface of the earth. Nagaraj (2018) brings together these two phenomena to demonstrate that otherwise random variation in the baseline availability of satellite maps resulted in upstream exploration-oriented firms taking advantage of sizeable differences in the timing of new gold deposits around the world (relative to integrated mining companies). Thus, variation in satellite image quality constrains mapmakers when the quality is low or images are unavailable and enables mapmakers when the quality exceeds a certain threshold.

Organizations and Incentives

In contrast to a traditional product or service, mapmaking often involves more than maximizing the profitability of selling a given map. Instead, it serves broader purposes of the organizations that fund the production and the cartographers that design that map. As information goods that involve the selective inclusion and exclusion of particular pieces of data, maps are often produced as a means to an end. For example, maps produced by Disney are given out for free, but are meant to stimulate demand for Disney-owned properties and attractions. In fact, as shown in Figure 7, the official Disney World map showing hotels in the area simply excludes a major state highway abutting its western edge, and represents the mixed residential

Figure 7
Maps of the Disney World Area in Orlando, FL, by Disney (Left) and Google Maps (Right)



Source: Disney map: https://www.wdwinfo.com/resortmaps/propertymap.htm. Google map: https://www.google.com/maps/@28.3855756,-81.5768293,13z.

areas adjoining the property (including non-Disney resort properties) as pristine wilderness even when alternate maps (such as Google) provide a more unbiased look. This map provides a very specific view of the Orlando area, aiming to maximize the engagement that Disney visitors have with theme park properties. The commercial goals of the sponsoring organization therefore have an important role to play in shaping the nature of maps.

A similar logic applies even when the mapmaking organization has noncommercial goals. A particularly striking example of the impact of commercial and nonprofit orientations of mapping can be seen in the mapping of refugee camps in areas such as Jordan, Nigeria, or the Gaza Strip produced over the past decade. In most areas of the world, and certainly in most locations with high levels of commercial activity, the for-profit Google Maps offers more or equally granular and detailed maps than open-source projects such as OpenStreetMap. However, as shown in Figure 8, the advantage turns to the nonprofit OpenStreetMap when one examines the establishment of high-quality maps and their dynamic updating for refugee camps (Palen et al. 2015). The prosocial motivations of OpenStreetMap volunteers have important implications for the maps that they produce.

Even for organizations with broadly similar objectives, differences in how they hope to achieve those objectives can result in significant alterations in map design. During the Cold War, Russian mapmakers made thousands of highly detailed 1:50,000 scale maps of many regions around the world, while the US military rarely made maps more detailed than 1:250,000, and those only covered areas of high strategic interest (Davies, Kent, and Risen 2017). These differences in mapping reflected differences in Cold War military strategies. Whereas the Soviet Union was focused on tank power and therefore required highly detailed cartographic maps

Figure 8
Maps for the Zaatari Refugee Camp, Jordan, on Google Maps (Left) Compared to OpenStreetMap (Right)

Source: Panel A: https://goo.gl/maps/pSwb8obFLTJTD2sM8. Panel B: https://www.openstreetmap.org/#map=15/32.2925/36.3215.

at a high level of resolution, the United States emphasized the importance of air power and required maps that covered a greater degree of terrain at a lower level of resolution.

It is important to emphasize that while our discussion has primarily focused on economic and prosocial incentives, maps are cultural and artistic products, and cartographers have long valued their artistic independence, demonstrating originality through design and aesthetics. One particularly salient example comes from the justly (in)famous New York subway map designed by Massimo Vignelli. Designed along modernist principles, this map prioritized a simple and clean look over accuracy; all routes ran at 45- or 90-degree angles, and Central Park was reconfigured as a square rather than a rectangle (Vignelli, Charysyn, and Noorda 1972). The uproar over its introduction ultimately led to a more traditional and informational representation, but this map has remained a favorite of modernist design critics to this day (Rawsthorn 2012). This simple but extreme example shows us that while map producers design maps according to their own idiosyncratic incentives, map users often need to rely on the information in the map without reference to how the underlying terrain has been distorted by those incentives.

Finally, while the factors of cost, demand, technology, competition, intellectual property, and organizations provide key shifters to the nature of maps, it is important to note that mapmaking is an endogenous and complex knowledge accumulation process. New maps build on preexisting ones, which are themselves shaped by these factors. The central feature that old maps influence newer ones creates path dependence in mapmaking that can lead to new information disseminating quickly across maps, but which could also cause large errors and inaccuracies to propagate

Figure 9
A French Map Depicting Baja California as an Island c. 1677

Source: Pierre Duval. "Carte Vniverselle du Monde Avec de nouvelles Observations: Amerique Septemtrionale." 1677. https://exhibits.stanford.edu/california-as-an-island/catalog/cb303zr7917.

for decades. The canonical example of this problem comes from the well-known case of California being depicted as an island on European maps throughout the seventeenth and into the eighteenth century. A Spanish expedition as early as 1539 (including many others) indicated that Baja California was a peninsula, and European maps initially represented it as such. However, starting in the early 1600s, most European maps depicted California as an island, as seen in Figure 9. Historians suggest that incorrect stories of Sir Francis Drake's travels in the Pacific in 1578 led to mapmakers across the European continent to make this error that was ultimately propagated across European maps for over 250 years (Polk 1995). Such path dependence creates strong linkages between newer and preexisting maps, and the five factors we highlight (cost, demand, innovation, competition, and organizations) can strengthen or weaken this link in important ways. While a full discussion of path dependence is beyond the purview of this essay, competition likely plays an important role. For example, there is likely to be a lower diversity in mapping representations when government or open maps are available to copy as opposed to more competitive settings where each provider must make maps from scratch, which would limit path dependence.

Concluding Thoughts

Our analysis has focused on the distinctive economic properties of maps. The features of territories that are mapped and those that are not are endogenously

shaped by the incentives and preferences of mapmakers. This area of inquiry is quite nascent and several theoretical, empirical, and policy challenges remain open.

First, on a theoretical level, we understand little about the equilibrium properties of maps. Why do maps that ostensibly have similar goals look different from one another in different settings and contexts (say, subway maps versus automobile maps)? Which mapping representations are more likely to succeed or fail? How do the factors that shape mapmaking interact and produce the maps that we see and use? While our framework is focused squarely on the agency of the mapmaker in shaping maps, how do users, data-providers, and policymakers shape the incentives of mapmakers through their own strategic interventions? Addressing such questions would help us clarify the relationship between social and private returns to mapmaking and identify industries and contexts where the two are likely to diverge.

Second, there are empirical challenges to consider when measuring the effects of different maps and mapmaking regimes on economic outcomes. In order to measure what was and was not included on a given map, we need a measure of ground truth distinct from the mapmaking project under study. For example, in order to examine which restaurants were not included on Yelp maps, Luca, Nagaraj and Subramani (2019) compared Yelp listings with administrative data from tax records. In many cases, a clean comparison is hard to achieve, especially when a mapping program includes features of the terrain that are uniquely captured in that map but not elsewhere. We need more empirical strategies to help provide a general methodology for work that tries to uncover the economic implications of endogenous variations in mapping.

Third, there are several open policy questions in this area. How can we systematically incorporate the idea that maps and geographic information not only describe geographies but also provide unique and (in our opinion) underutilized tools to shape geography? For example, consider the recently released Startup Cartography Project (Andrews et al. 2017) that provides highly granular maps of high-potential entrepreneurial activity in the United States. These maps not only describe the state of American entrepreneurship (Guzman and Stern 2016), but also provide policy guidance to startups on where they should locate and to policymakers on where they should focus their efforts. Similarly, intergenerational mobility maps provided by Chetty et al. (2014) are being used by policymakers to guide the allocation of resources across geographies. How should such maps be designed to maximize social returns? How can maps be incorporated into a policy toolkit, and what are some general processes of map design that maximize social welfare?

Finally, while we focused our attention on geographic maps in this essay, our work has broader implications for maps of non-geographic spaces as well. For example, some work in economics has studied the development of the human genome map (Williams 2013; Jayaraj and Gittelman 2018; Kao 2019) and its role in shaping the direction of pharmaceutical innovation. Similarly, planetary and space maps of various kinds are important in the development of astronomical and astrophysical models. Industry maps and the idea of "mental mapping" are also commonly used metaphors

in business (Puranam and Swamy 2016). Our basic framework that separates mapping representations from the terrain and focuses on the mapmaker's endogenous selection process could be equally applicable in these scenarios.

■ *We are grateful to Enrico Moretti, Gordon Hanson, Heidi Williams, and Timothy Taylor for their advice and guidance in developing this essay. We would also like to thank R.J. Andrews, Laura Bliss, Jorge Guzman, and Aruna Ranganathan for providing insightful comments on a previous version of this essay as well as Leon Ming and, especially, Melissa Staha for excellent research assistance.*

References

Aaronson, Daniel, Daniel A. Hartley, and Bhashkar Mazumder. 2017. "The Effects of the 1930s Holc 'Redlining' Maps." FRB of Chicago Working Paper WP-2017-12. https://ssrn.com/abstract=3038733.

AlphaBeta. 2017. "The Economic Impact of Geospatial Services: How Consumers, Businesses, and Society Benefit from Location-based Information." https://www.alphabeta.com/wp-content/uploads/2017/09/GeoSpatial-Report_Sept-2017.pdf.

Andrews, Raymond J., Catherine Fazio, Jorge Guzman, and Scott Stern. 2017. "The Startup Cartography Project: A Map of Entrepreneurial Quality and Quantity in the United States across Time and Location." https://static1.squarespace.com/static/5963ccedebbd1a0ffdb5ae00/t/596624123e00be701a3737d3/1499867278836/Andrews+Fazio+Guzman+Stern+%E2%80%94+Startup+Cartography+Project+%E2%80%94+EARLY+DRAFT.pdf.

Ash, Eric H. 2007. "Navigation Techniques and Practice in the Renaissance." In *The History of Cartography*, Vol. 3, edited by David Woodward, 509–27. Chicago: The University of Chicago Press.

Baragwanath, Kathryn, Ran Goldblatt, Gordon Hanson, and Amit K. Khandelwal. 2019. "Detecting Urban Markets with Satellite Imagery: An Application to India." *Journal of Urban Economics*.

Barrington-Leigh, Christopher, and Adam Millard-Ball. 2017. "The World's User-Generated Road Map is More than 80% Complete." *PLoS ONE* 12 (8): e0180698. https://doi.org/10.1371/journal.pone.0180698.

Barrow, Ian J. 2004. *Making History, Drawing Territory: British Mapping in India, c. 1756–1905*. New Delhi: Oxford University Press.

Berry, Steven T., and Joel Waldfogel. 1999. "Public Radio in the United States: Does It Correct Market Failure or Cannibalize Commercial Stations?" *Journal of Public Economics* 71 (2): 189–211.

Bloomberg. 2018. "Nobody Wants to Let Google Win the War for Maps All Over Again." *Fortune*, February 21. https://fortune.com/2018/02/21/google-waymo-mapping-software/.

Boston Consulting Group (BCG). 2012. *Putting the U.S. Geospatial Services Industry on the Map*. https://media.nationalgeographic.org/assets/file/BostonConsultingGroup_US_FullReport.pdf.

Bridle, James. 2012. "Trap Streets: The Road Not Taken." *Cabinet Magazine*, Fall. http://www.cabinetmagazine.org/issues/47/bridle.php.

Brynjolfsson, Erik, Avinash Collis, and Felix Eggers. 2019. "Using Massive Online Choice Experiments to Measure Changes in Well-Being." *Proceedings of the National Academy of Science* 116 (15): 7250–55.

BuiltWith. 2019. "Mapping Usage Distribution in the Top 1 Million Sites." Sydney, Australia: BuiltWith. https://trends.builtwith.com/mapping.

Casaburi, Lorenzo, and Ugo Troiano. 2016. "Ghost-House Busters: The Electoral Response to a Large Anti-tax Evasion Program." *Quarterly Journal of Economics* 131 (1): 273–314.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where Is the Land of

Opportunity? The Geography of Intergenerational Mobility in the United States." *Quarterly Journal of Economics* 129 (4): 1553–1623.

Clement, J. 2018. "Most Popular Mapping Apps in the United States as of April 2018, by Monthly Users." *Statista*, May 29. https://www.statista.com/statistics/865413/most-popular-us-mapping-apps-ranked-by-audience/.

Croft, Thomas A. 1978. "Nighttime Images of the Earth from Space." *Scientific American* 239 (1): 86–101.

Daley, Suzanne. 2010. "Greek Wealth is Everywhere but Tax Forms." *New York Times*, May 1. https://www.nytimes.com/2010/05/02/world/europe/02evasion.html.

Daum, Meghan. 2015. "Letter of Recommendation: The Thomas Guide to Los Angeles." *New York Times*, April 3. https://www.nytimes.com/2015/04/05/magazine/letter-of-recommendation-the-thomas-guide-to-los-angeles.html.

Davies, Arthur. 1977. "Behaim, Martellus and Columbus." *Geographical Journal* 143 (3): 451–59. https://www.jstor.org/stable/634713?seq=1.

Davies, John, Alexander J. Kent, and James Risen. 2017. *The Red Atlas: How the Soviet Union Secretly Mapped the World*. Chicago: University of Chicago Press.

Dell, Melissa. 2010. "The Persistent Effects of Peru's Mining Mita." *Econometrica* 78 (6): 1863–1903. http://www.econometricsociety.org/tocs.asp.

DiBiase, David. 2008. *Nature of Geographic Information: An Open Geospatial Textbook*. State College, PA: Pennsylvania State University. https://opentextbc.ca/natureofgeographicinformation/.

Donaldson, Dave, and Adam Storeygard. 2016. "The View from Above: Applications of Satellite Data in Economics." *Journal of Economic Perspectives* 30 (4): 171–98.

ESRI. 2016. "State of North Carolina MPA Price List E412F-1Q2017." April 11. https://files.nc.gov/ncdit/documents/files/E412F-1Q2017-Price-list-Time-Material-Costs.pdf.

European Global Navigation Satellite Systems Agency. 2017. *GNSS Market Report*. Issue 5. https://www.gsa.europa.eu/system/files/reports/gnss_mr_2017.pdf.

Geospatial Media and Communications. 2019. *GeoBuiz 2019 Report: Geospatial Industry Outlook and Readiness Index*. https://geobuiz.com/geobuiz-report-2019.

Glaeser, Edward L., Hedi D. Kallal, José A. Scheinkman, and Andrei Shleifer. 1992. "Growth in Cities." *Journal of Political Economy* 100 (6): 1126–152.

GPS.gov. 2019. "Fiscal Year 2019 Program Funding." https://www.gps.gov/policy/funding/2019/.

Guo, Zhan. 2011. "Mind the Map! The Impact of Transit Maps on Path Choice in Public Transit." *Transportation Research Part A: Policy and Practice* 45 (7): 625–39.

Guzman, Jorge, and Scott Stern. 2016. "The State of American Entrepreneurship: New Estimates of the Quantity and Quality of Entrepreneurship for 15 US States, 1988–2014." NBER Working Paper 22095.

Harley, John Brian. 1989. "Deconstructing the Map." *Cartographica: The International Journal for Geographic Information and Geovisualization* 26 (2): 1–20.

Helft, Miguel. 2016. "The Godfather of Digital Maps." *Forbes*, February 10. https://www.forbes.com/sites/miguelhelft/2016/02/10/the-godfather-of-digital-maps/#357afcdf4da9.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102 (2): 994–1028.

Hook, Leslie. 2016. "Uber to Pour $500m into Global Mapping Project." *Financial Times*, July 31. https://www.ft.com/content/e0dfa45e-5522-11e6-befd-2fc0c26b3c60.

Jacobs, Frank. 2014. "Agloe: How a Completely Made Up New York Town Became Real." *Big Think*, February 12. https://bigthink.com/strange-maps/643-agloe-the-paper-town-stronger-than-fiction.

Jayaraj, Sebastian, and Michelle Gittelman. 2018. "Scientific Maps and Innovation: Impact of the Human Genome Map on Drug Discovery." http://docplayer.net/129325399-Scientific-maps-and-innovation-impact-of-the-human-genome-map-on-drug-discovery-sebastian-jayaraj-a-dissertation-submitted-to-the.html.

Kao, Jennifer. 2019. "Charted Territory: Evidence from Mapping the Cancer Genome and R&D Decisions in the Pharmaceutical Industry." https://scholar.harvard.edu/files/kao/files/kao_jmp_cancer_pharma.pdf.

Katona, Zsolt, Marcus O. Painter, Panos N. Patatoukas, and Jieyin Zeng. 2018. "On the Capital Market Consequences of Alternative Data: Evidence from Outer Space." 9th Miami Behavioral Finance Conference 2018.

Kershner, Jim. 2013. "Britain and the United States Agree on the 49th Parallel as the Main Pacific Northwest Boundary in the Treaty of Oregon on June 15, 1846." *HistoryLink*, July 31. https://historylink.org/File/5247.

Korzybski, Alfred. 1933. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*.

New York: The International Non-Aristotelian Library Publishing Company.

Lambert, Fred. 2017. "Tesla is Updating its Maps and Navigation with Open Source Mapping Platforms." *Electrek*, July 3. https://electrek.co/2017/07/03/tesla-map-navigation-open-source-platforms/.

Landes, William M., and Richard A. Posner. 1989. "An Economic Analysis of Copyright Law." *Journal of Legal Studies* 18 (2): 325–63.

Larcom, Shaun, Ferdinand Rauch, and Tim Willems. 2017. "The Benefits of Forced Experimentation: Striking Evidence from the London Underground Network." *Quarterly Journal of Economics* 132 (4): 2019–55.

Lewis, Michael. 2018. *The Coming Storm*. Audible Audio: Audible Original.

Luca, Michael, Abhishek Nagaraj, and Gauri Subramani. 2019. "Getting on the Map: The Impact of Online Listings on Business Performance." Unpublished.

Lunden, Ingrid. 2013. "Google Bought Waze for $1.1B, Giving a Social Data Boost to its Mapping Business." *TechCrunch*, June 11. https://techcrunch.com/2013/06/11/its-official-google-buys-waze-giving-a-social-data-boost-to-its-location-and-mapping-business/.

Maher, William J. 2002. "Copyright Term, Retrospective Extension, and the Copyright Law of 1790 in Historical Context." *Journal of the Copyright Society of the USA* 49 (4): 1021–39.

Mankiw, N. Gregory, and Michael D. Whinston. 1986. "Free Entry and Social Inefficiency." *RAND Journal of Economics* 17 (1): 48–58.

Maxar Technologies. 2019. "Maxar Technologies Awarded Four-Year Global EGD Contract by the U.S. Government for On-Demand Access to Mission-Ready Satellite Imagery." *Sensor and Systems*, August 28. https://sensorsandsystems.com/maxar-technologies-awarded-four-year-global-egd-contract-by-the-u-s-government-for-on-demand-access-to-mission-ready-satellite-imagery/.

MDA Corporation. 2017. "MDA to Acquire DigitalGlobe, Creating Industry Leader in End-to-End Space Systems, Earth Imagery, and Geospatial Solutions." February 24. https://mdacorporation.com/news/pr/pr2017022402.html.

Michel-Kerjan, Erwann O. 2010. "Catastrophe Economics: The National Flood Insurance Program." *Journal of Economic Perspectives* 24 (4): 165–86.

Milgrom, Paul R., Douglass C. North, and Barry R. Weingast. 1990. "The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs." *Economics and Politics* 2 (1): 1–23.

Monmonier, Mark, and H.J. de Blij. 1996. *How to Lie with Maps*. 2nd ed. Chicago: University of Chicago Press.

Moretti, Enrico. 2012. *The New Geography of Jobs*. Boston: Houghton Mifflin Harcourt.

Nagaraj, Abhishek. 2018. "The Private Impact of Public Information—Landsat Satellite Maps and Gold Exploration." http://abhishekn.com/files/nagaraj_landsat_oct2018.pdf.

Nagaraj, Abhishek, and Henning Piezunka. 2018. "Deterring the New, Motivating the Established—The Divergent Effect of Platform Competition on Member Contributions in Digital Mapping Communities." http://abhishekn.com/files/nagaraj_piezunka_competition.pdf.

Nagaraj, Abhishek, Esther Shears, and Mathijs de Vaan. 2018. "Does Data Access Democratize Science?" Unpublished.

Naik, Nikhil, Ramesh Raskar, and César A. Hidalgo. 2016. "Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance." *American Economic Review* 106 (5): 128–32.

O'Bierne Justin. 2016. "Cartography Comparison." https://www.justinobeirne.com/cartography-comparison (accessed December 1, 2019).

O'Connor, Claire. 2018. *IBISWorld Industry Report 54137: Surveying and Mapping Services in the US*. Los Angeles: IBISWorld.

Palen, Leysia, Robert Soden, T. Jennings Anderson, and Mario Barrenechea. 2015. "Success and Scale in a Data-Producing Organization: The Socio-Technical Evolution of OpenStreetMap in Response to Humanitarian Events." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4113–22. Association for Computing Machinery.

Politico. 2016. "2016 Oklahoma Presidential Election Results." December 13. https://www.politico.com/2016-election/results/map/president/oklahoma/.

Polk, Dora Beale. 1995. *The Island of California: A History of the Myth*. Lincoln, NE: University of Nebraska Press.

Popper, Ben. 2017. "Google Announces Over 2 Billion Monthly Active Devices on Android." *The Verge*, May 17. https://www.theverge.com/2017/5/17/15654454/android-reaches-2-billion-monthly-active-users.

Puranam, Phanish, and Murali Swamy. 2016. "How Initial Representations Shape Coupled Learning Processes." *Organization Science* 27 (2): 323–35.

Rawsthorn, Alice. 2012. "The Subway Map That Rattled New Yorkers" *New York Times*, August 5. https://www.nytimes.com/2012/08/06/arts/design/the-subway-map-that-rattled-new-yorkers.html.

Robinson, Arthur H., Joel L. Morrison, Phillip C. Muehrcke, A. Jon Kimerling, and Stephen C. Guptill. 1995. *Elements of Cartography*. 6th ed. New York: Wiley.

Schaal, Dennis. 2019. "Google Maps Poised to Be an $11 Billion Business in 4 Years." *Skift*, August 30. https://skift.com/2019/08/30/google-maps-poised-to-be-an-11-billion-business-in-4-years/.

Schulten, Susan. 2012. "Mismapping the Peninsula." *New York Times,* April 20. https://opinionator.blogs.nytimes.com/2012/04/20/mismapping-the-peninsula/.

Scott, James C. 1999. *Seeing like a State: How Certain Schemes to Improve the Human Condition have Failed*. New Haven: Yale University Press.

Singh, Bajinder Pal. 2018. "When Disasters Strike, Satellites Come Calling." *ReliefWeb*, October 22. https://reliefweb.int/report/indonesia/when-disasters-strike-satellites-come-calling.

TomTom. 2019. "Q4 and FY 2018 Results." TomTom Investor Relations, February 6. https://corporate.tomtom.com/static-files/65db0677-d017-4c06-9b0b-54c0bc4329b0.

Uber. 2019. "Form S-1 Registration Statement." Securities and Exchange Commission. Filed on April 11. https://www.sec.gov/Archives/edgar/data/1543151/000119312519103850/d647752ds1.htm.

Upshot Staff. 2018. "Political Bubbles and Hidden Diversity: Highlights from a Very Detailed Map of the 2016 Election." *New York Times*, July 25. https://www.nytimes.com/interactive/2018/07/25/upshot/precinct-map-highlights.html.

V1 Media. 2019. "Maxar Technologies Awarded Four-Year Global EGD Contract by the U.S. Government for On-Demand Access to Mission-Ready Satellite Imagery." *Sensor and Systems*, August 28. https://sensorsandsystems.com/maxar-technologies-awarded-four-year-global-egd-contract-by-the-u-s-government-for-on-demand-access-to-mission-ready-satellite-imagery/.

Vietor, Alexander O. 1963. "A Pre-Columbian Map of the World, Circa 1489." *Imago Mundi* 17: 95–96.

Vignelli, Massimo, Joan Charysyn, and Bob Noorda. 1972. "New York Subway Map." *MoMA*. https://www.moma.org/collection/works/89300.

Wang, Jules. 2019. "Google Maps Revenue Expected to Increase, Following API Price Hikes and Planned Ads." *Android Police*, April 11. https://www.androidpolice.com/2019/04/11/google-maps-revenue-expected-to-increase-following-api-price-hikes-and-planned-ads/.

Williams, Heidi L. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *Journal of Political Economy* 121 (1): 1–27.

Wood, Mark. 2014. "The Island of California." *Pomona College Magazine*, March 11. http://magazine.pomona.edu/2014/spring/the-island-of-california/.

World Bank. 2010. *2010 Haiti Earthquake Final Report.* https://www.gfdrr.org/sites/default/files/publication/2010haitiearthquakepost-disasterbuildingdamageassessment.pdf.

Wright, Gavin. 1990. "The Origins of American Industrial Success, 1879–1940." *American Economic Review* 80 (4): 651–68.

Wright, Robin. 2016. "How the Curse of Sykes-Picot Still Haunts the Middle East." *The New Yorker,* April 30. https://www.newyorker.com/news/news-desk/how-the-curse-of-sykes-picot-still-haunts-the-middle-east.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 222–239

Emi Nakamura: 2019 John Bates Clark Medalist

Janice Eberly and Michael Woodford

Emi Nakamura, winner of the 2019 John Bates Clark Medal, is an empirical macroeconomist. She has made a signature contribution to the field by integrating microeconomic and macroeconomic theory with data to increase our understanding of some of the most consequential, challenging, and long-standing questions in macroeconomics. Emi's distinctive approach displays a sophisticated understanding of alternative theoretical models of macroeconomic phenomena and then turns to both their unique micro implications and aggregate consequences to distinguish between models. In the past, most prior work on these big macroeconomic questions has been built on quarterly, aggregate time series for the post-World War II period. In contrast, Emi analyzes macro questions by considering implications that arise in more disaggregated, or higher frequency data, or extending over a longer historical period. Her empirical work requires painstaking analysis of data sources not previously exploited, and she has been notably creative in developing and using new sources of data. By bringing together insightful modeling and new data resources, she has managed to isolate variation in micro data that is more credible for drawing causal inferences. Moreover, she can then relate these results to earlier estimation approaches, interpreting existing evidence in light of new methods and models.

■ *Janice Eberly is James R. and Helen D. Russell Professor of Finance, Kellogg School of Management, Northwestern University, Evanston, Illinois. Michael Woodford is John Bates Clark Professor of Political Economy, Columbia University, New York City, New York. Their email addresses are eberly@kellogg.northwestern.edu and mw2230@columbia.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.1.222.

Emi Nakamura

Emi's exposure to economics began early in life. Her grandfather, Guy Orcutt, was a distinguished econometrician (Watts 1991). Both of her parents, Alice and Masao Nakamura, were academic economists; her mother, Alice Orcutt Nakamura, is a past President of the Canadian Economic Association. In addition to an early exposure to economic ideas, Emi credits her parents with instilling in her "a deep sense of the importance of testing theories empirically" (Ng 2015). Emi attended academic conferences with her mother and began taking economics classes at the University of British Columbia as a high school student. She credits one of these early classes, a master's class on economic measurement and index number theory taught by Erwin Diewert, with making an early mark in her drive for clarity in measurement. In a similar vein, Emi watched the film "The Race for the Double Helix" about the discovery of the structure of DNA with her parents. They emphasized the role of the empiricist Rosalind Franklin and the notion that "there is nothing worse than a wrong fact."

During her undergraduate studies at Princeton, she took many of the graduate classes in economics. This included Bo Honore's graduate course in econometrics, where she pressed forward her interest in measurement and estimation, and also met her future husband and frequent coauthor, Jón Steinsson, who was also a Princeton undergraduate at the time. Emi's interest in macroeconomics was piqued at Princeton under the guidance of her undergraduate advisor, Mike Woodford. She particularly remembers writing to him with a question on the real business cycle model over the winter break and receiving a detailed response on Christmas Day. A revised version of her Princeton senior thesis was published as [5], shown in Table 1.

Table 1
Selected Publications of Emi Nakamura

1. "Cost Pass-Through in the U.S. Coffee Industry" (with Ephraim Leibtag, Alice Nakamura, and Dawit Zerom). 2007. Economic Research Report Number 38.
2. "Layoffs and Lemons over the Business Cycle." 2008. *Economics Letters* 99 (1): 55–58.
3. "Pass-Through in Retail and Wholesale." 2008. *American Economic Review: Papers & Proceedings* 98 (2): 430–37.
4. "Five Facts about Prices: A Reevaluation of Menu Cost Models" (with Jón Steinsson). 2008. *Quarterly Journal of Economics* 123 (4): 1415–64.
5. "Deconstructing the Success of Real Business Cycles." 2009. *Economic Inquiry* 47 (4): 739–53.
6. "Accounting for Incomplete Pass-Through" (with Dawit Zerom). 2010. *Review of Economic Studies* 77 (3): 1192–1230.
7. "Monetary Non-Neutrality in a Multi-Sector Menu Cost Model" (with Jón Steinsson). 2010. *Quarterly Journal of Economics* 125 (3): 961–1013.
8. "Price Dynamics, Retail Chains and Inflation Measurement" (with Alice O. Nakamura and Leonard I. Nakamura). 2011. *Journal of Econometrics* 161 (1): 47–55.
9. "Price Setting in Forward-Looking Customer Markets" (with Jón Steinsson). 2011. *Journal of Monetary Economics* 58 (3): 220–33.
10. "Lost in Transit: Product Replacement Bias and Pricing to Market" (with Jón Steinsson). 2012. *American Economic Review* 102 (7): 3277–3316.
11. "Crises and Recoveries in an Empirical Model of Consumption Disasters" (with Jón Steinsson, Robert Barro, and José Ursúa). 2013. *American Economic Journal: Macroeconomics* 5 (3): 35–74.
12. "Price Rigidity: Microeconomic Evidence and Macroeconomic Implications" (with Jón Steinsson). 2013. *Annual Review of Economics* 5: 133–63.
13. "Fiscal Stimulus in a Monetary Union: Evidence from US Regions" (with Jón Steinsson). 2014. *American Economic Review* 104 (3): 753–92.
14. "Are Chinese Growth and Inflation Too Smooth? Evidence from Engel Curves" (with Jón Steinsson and Miao Liu). 2016. *American Economic Journal: Macroeconomics* 8 (3): 113–44.
15. "The Power of Forward Guidance Revisited" (with Alisdair McKay and Jón Steinsson). 2016. *American Economic Review* 106 (10): 3133–58.
16. "Growth-Rate and Uncertainty Shocks in Consumption: Cross-Country Evidence" (with Dmitriy Sergeyev and Jón Steinsson). 2017. *American Economic Journal: Macroeconomics* 9 (1): 1–39.
17. "Informational Rigidities and the Stickiness of Temporary Sales" (with Eric Anderson, Benjamin A. Malin, Duncan Simester, and Jón Steinsson). 2017. *Journal of Monetary Economics* 90: 64–83.
18. "Identification in Macroeconomics" (with Jón Steinsson). 2018. *Journal of Economic Perspectives* 32 (3): 59–86.
19. "High-Frequency Identification of Monetary Non-Neutrality: The Information Effect" (with Jón Steinsson). 2018. *Quarterly Journal of Economics* 133 (3): 1283–1330.
20. "The Elusive Costs of Inflation: Price Dispersion during the U.S. Great Inflation" (with Jón Steinsson, Patrick Sun, and Daniel Villar). 2018. *Quarterly Journal of Economics* 133 (4): 1933–80.

Emi went on to graduate school to study economics at Harvard, where she took full advantage of the rich curriculum and varied methodological offerings. Taking Caroline Hoxby's empirical labor course back-to-back with Ariel Pakes's structural industrial organization course turned out to be a fruitful pairing, as Emi recalls feeling challenged to integrate Hoxby's description of the "revolution in identification" with the sophisticated structural models from Pakes's class. Emi's research collaboration with Jón began during graduate school, and she received her PhD from Harvard in 2007.

Emi began her career at Columbia University, where she held joint appointments in the Department of Economics and the Graduate School of Business. At Columbia, she received tenure in 2013 and was promoted to full professor in 2017. Since 2018, she has been the Chancellor's Professor of Economics at the University of California, Berkeley.

Emi is often asked what it is like to work so closely with her husband, both in research and in child-rearing. They both describe the communication necessary to succeed in all these roles as complementary. And of course, their willingness to apply principles of economic efficiency helps, as well. Recognizing that they do not have to do everything themselves liberates time for activities at which they excel (like research) and which they especially value (like child-rearing). Their family choices were described as "out-sourcing" by the *New York Times* (Rampell 2013), but many working parents will recognize such trade-offs and survival skills.

Of course, the American Economic Association is not the first to recognize Emi's promise and accomplishments, which include a CAREER Award from the NSF (2011), a Sloan Research Fellowship (2014), the Elaine Bennett Research Prize from the AEA (2014), and being named a member of "Generation Next: Top 25 Economists Under 45" by the IMF (2014). As her work has become more influential, she has in turn influenced other scholars by taking on leadership roles in the economics profession: for example, she serves as a co-editor of the *American Economic Review* and as co-director of the NBER Program on Monetary Economics. She also serves on the Panel of Economic Advisers for the Congressional Budget Office, the AEA Committee on National Statistics, and the Technical Advisory Committee of the Bureau of Labor Statistics. These appointments testify to the role she has quickly gained in the profession as an expert on issues relating to data construction and use.

The breadth of Emi's research agenda is apparent in the five main topics we discuss here: 1) Models of Price Adjustment, 2) Models of Pass-Through of Costs to Prices, 3) Empirical Studies of Asset Pricing, 4) Empirical Studies of Fiscal Stimulus, and 5) The Effects of Monetary Policy. We refer to her papers by number, as enumerated in Table 1.

Models of Price Adjustment

Emi is arguably best known for her work on the nature and consequences of price rigidity. Her research on this general topic encompasses both theoretical and

empirical work on a variety of models of price adjustment, and it considers implications of price rigidity for both domestic and open-economy issues.

Her most widely known paper is "Five Facts about Prices: A Reevaluation of Menu Cost Models" [4], with Jón Steinsson. This paper is a key reference in one of the more important recent developments in monetary economics, which is studying price adjustment by looking at changes in individual prices, rather than just using aggregate price indices. Measures of the average time that prices of individual goods remain unchanged have long been an important source of evidence for price rigidity. However, until very recently, most evidence of this kind came from detailed studies of a very small number of markets. Availability of new datasets that allow changes in the prices of a very large number of goods to be tracked simultaneously has radically transformed this literature, and Emi and Jón's careful work in [4] has been one of the most influential contributions.

Emi and Jón study the data from the Bureau of Labor Statistics on individual prices used to construct the Consumer and Producer Price Indexes. They document a variety of facts about changes in individual prices that can be compared to the implications of a popular theoretical model of price adjustment, the "menu cost" model. For example, while past studies using other sources had concluded that the median time between price changes in the US economy was a large fraction of a year, the first work using the BLS micro data by Bils and Klenow (2004) had argued that prices actually changed much more frequently, with a median duration of prices only a little over four months.

However, Bils and Klenow (2004) used an extract from the Bureau of Labor Statistics micro dataset, for the period between 1995 and 1997. In [4], Emi and Jón obtained access to the BLS micro data containing all of the price observations collected for the period from 1988 to 2005. Emi and Jón show that conclusions about the frequency of price changes depend on the method used to distinguish sales from changes in "regular prices." They find both that changes in "regular prices" occur much less often than price changes that include sales (they find a median duration of 8–11 months for "regular prices," depending on the precise method used to classify price changes), and that producer prices (for which there is less of a need to filter out "sales") also change quite infrequently. This paper suggests that the microeconomic evidence for substantial "stickiness" of individual prices is considerably stronger than Bils and Klenow had implied.

In addition, the paper [4] documents several features of the data on individual price changes that can be used to test popular models of price adjustment. Emi and Jón stress two features of the data in particular that are contrary to the predictions of popular "menu cost" models of price adjustment: clear seasonality in the frequency of price adjustments and the failure of the likelihood of price changes to increase with the amount of time that has passed since the last change in price. In contrast, "menu cost" models emphasize that changing prices has a cost, and so if the existing nominal price becomes less appropriate over time—perhaps because of inflation or changes in cost conditions—the price adjustments will happen only after a lag and will often involve substantial discrete jumps.

The ability of a "menu cost" model to account for the quantitative characteristics of the micro data on price changes is considered further in Emi's paper "Monetary Non-Neutrality in a Multi-Sector Menu Cost Model" [7]. Prior numerical analyses of the implications of menu cost models, such as the very influential paper by Golosov and Lucas (2007), had used a one-sector model which assumed that all goods in the economy were subject to menu costs of the same size, in addition to being produced with the same technology, and so on. In this approach, the parameters common to all goods were assigned numerical values to match statistics for the set of all price changes, such as the overall frequency of change in prices and the average absolute size of price changes. But one of the facts documented by Emi and Jón in [4] is that there is tremendous heterogeneity across sectors of the US economy in the frequency of (nonsale) price changes.

In [7], Emi and Jón calibrate a multi-sector menu cost model to match the distribution across sectors of both the frequency of price changes and the average size of price changes. They find that the real effects of a monetary disturbance are three times as large in their multi-sector model as in a one-sector model, like that of Golosov and Lucas (2007). Indeed, whereas Golosov and Lucas argue that price rigidity is not an empirically plausible explanation for the observed effects of monetary disturbances in their one-sector model, Emi and Jón show that their calibrated multi-sector model (with nominal shocks of the magnitude observed for the US economy) predicts output fluctuations that would account for nearly one-quarter of the US business cycle. This magnitude would be roughly in line with the fraction of GDP variability that is attributed to monetary disturbances in atheoretical vector-autoregression studies. The emphasis of [7] on the importance of taking sectoral heterogeneity into account when parameterizing the degree of price stickiness has been highly influential.

Emi and Jón have also addressed the open-economy implications of alternative models of price-setting. Their paper "Lost in Transit: Product Replacement Bias and Pricing to Market" [10] looks at microeconomic data on individual price changes to reassess an important issue in open-economy macroeconomics, which is the extent to which exchange-rate changes are "passed through" to changes in the prices of US imports and exports. Previous literature had suggested that the relative prices of US imports change by only 0.2 to 0.4 percent in the case of a 1 percent change in the exchange rate, while the relative price of US exports changes by nearly 1 percent. This incomplete adjustment of import prices (even after substantial periods of time) is often taken as evidence of "pricing to market" by the foreign suppliers of US imports, whereas US exporters evidently "price to market" to a much lower extent.

However, Emi and Jón argue in [10] that conventional measures are seriously biased, owing to measurement errors created by price rigidity and relatively frequent product replacement. They show that as a result of these factors, about 45 percent of the individual price series used to construct the US import and export price series have no price changes at all, while roughly 70 percent have only two price changes or fewer over the time that price is measured. Emi and Jón argue that a large number of price changes occur at the time of product replacements,

but are ignored in the construction of the indices—because changes in the index frequently reflect only price changes that occur in the case of a good whose characteristics have not also changed. They estimate the magnitude of the bias that this produces in measures of "pass-through" to be as large as a factor of two. When they correct for the bias, they find that relative import prices respond by 0.6–0.7 of the size of the change in the exchange rate, while the relative price of US exports responds by only 0.8 of the change in the exchange rate. Thus, their results suggest that there is much less difference in the behavior of US exporters and exporters to the US economy than is commonly believed. Again, the use of micro pricing data can shed light on the nature of price adjustment that was not obtainable by previous studies using aggregate price indices.

One of Emi's major research efforts in recent years has been a labor-intensive multi-year project of extending the BLS micro-level dataset on consumer prices back in time by more than a decade to 1977. This project required more than the usual amount of empirical resourcefulness, as Emi found the data on microfilm cartridges in old file cabinets at the Bureau of Labor Statistics. These cartridges were not readable with modern equipment, nor could they be taken out of the Bureau of Labor Statistics. Even when a machine could be retrofitted, the scans had to be done by the Bureau of Labor Statistics (on their budget and staff-time) to meet confidentiality and ethics requirements. Finally, the scans resulted in a million PDF files, which could not leave the Bureau of Labor Statistics for transcription. Emi and Jón worked with a developer to create an optical character recognition program of sufficient accuracy to convert the images to machine-readable data. One of many advantages is that the resulting extended database includes a period in the late 1970s and early 1980s when inflation was much higher and more volatile than it has been since 1988 and also a period of deep recession. There will be much more scope to study how patterns of price adjustment change in response to changing macroeconomic conditions—an issue of central importance for macroeconomic uses of models of price-setting.

A first (though likely not the last) important paper using this new dataset is "The Elusive Costs of Inflation: Price Dispersion during the U.S. Great Inflation" [20], written by Emi and Jón with Patrick Sun and Daniel Villar. The paper considers how the process of adjustment of firms' prices to changing market conditions differs in a higher inflation environment—a question that is important for assessing the welfare costs of higher inflation. They find that "regular" (nonsale) prices were adjusted more frequently in the earlier higher inflation part of their dataset and by about the amount that would be predicted by a model of optimal price adjustment taking into account a fixed "menu cost" of adjusting the firm's price. They conclude that in assessing the welfare costs expected to follow from a permanently higher rate of inflation, it is important to take into account the increased frequency of price adjustments that should be expected to occur.

The paper also seeks to measure the degree to which there is greater dispersion in the prices of similar products in a higher inflation environment. Some common models of price adjustment imply that price dispersion should rise in a high-inflation

setting, owing to staggering of the times at which different firms' prices happen to be reconsidered. However, measuring price dispersion is difficult because it can be hard to tell if different prices across firms might just reflect heterogeneity of the goods. For this reason, Emi and her coauthors take an indirect approach: they look at how the average size of price changes differs between high- and low-inflation periods and find that the average size of price increases, when they occur, is about the same (a 7 percent increase on average) in their pre-1988 sample as in their post-1988 sample. Thus, they argue that a higher inflation rate does not increase price dispersion.

This paper [20] is an important contribution along several dimensions: to policy debates about the costs of inflation, to our understanding of historical facts about price adjustment in the United States, and to the empirical basis for assessing the realism of alternative theoretical models of price-setting. It further cements Emi and Jón's reputations as preeminent experts on price dynamics and the empirical evidence for models of price-setting, as already indicated by their 2013 review article on the topic: "Price Rigidity: Microeconomic Evidence and Macroeconomic Implications" [12].

Emi and Jón have also made theoretical contributions to models of price-setting. As mentioned earlier, an important pattern that they observe about the micro price data is that many goods tend to have a "regular" price that changes infrequently, while various "sale" prices are also charged at times between occasions on which the "regular" price changes. In "Price Setting in Forward-Looking Customer Markets" [9], Emi and Jón offer a theoretical explanation for such a dynamic pattern of prices in the context of a dynamic model of price setting in the context of a "deep habits" model (Ravn, Schmitt-Grohé, and Uribe 2006). In this approach, the demand for goods that a firm faces depends not only on the current price the firm charges but also on past sales (because households have habit-forming preferences) and on expected future prices of the good (because households know they have habit-forming preferences and worry about becoming accustomed to consuming a good with high future prices). In this setting, firms have an incentive to use time-inconsistent pricing policies: that is, they want to promise low prices in the future to attract customers today, but once consumers have developed a habit of consuming this good, firms have an incentive to break their promises and start charging high prices.

Characterizing time-consistent pricing strategies in this environment is challenging, and Emi and Jón compare several possibilities in [9]. In several of these approaches, a situation arises of time-consistent price dynamics in which regular prices will appear to be nonresponsive to variations in the exogenous state and in which sales prices will involve varying discounts and hence can be interpreted as responding to the exogenous state of the economy. In one such approach, a firm's prices can be contingent on its past prices, using an equilibrium concept similar to the one that Chari and Kehoe (1990) refer to as "sustainable plans." In this case, Emi and Jón show that there exist time-consistent equilibria in which the price is unresponsive to the values of exogenous shocks in (for example) all even periods, while it is responsive to shocks in all odd periods.

Another interesting finding in this paper [9] arises when the firm has private information about marginal costs and the strength of demand. For this environment, Emi and Jón apply theoretical results from the work of Athey, Bagwell, and Sanchirico (2004) to show that the sustainable price that maximizes the value of the firm has the feature that there is a maximum price cap that the firm does not exceed, even if the exogenous state variables exceed a certain threshold. Again, price data from this economy would potentially be consistent with the pattern of a regular price that is visited frequently and with the observation of temporary sales with flexible prices. In an empirical analysis using the Dominick's Finer Foods database of supermarket prices, they find that the frequency of sale price adjustment is about eight times higher than the frequency of regular price adjustment, which they interpret as supporting the prediction of their theoretical model. This is an ambitious and highly creative paper on a challenging and important topic.

Models of Pass-Through of Costs to Prices

In work that bridges industrial organization, international trade, and macroeconomics, Emi has studied the pass-through of changes in costs to the prices that firms charge for their products.

One of Emi's first papers [1] studied the pass-through of foreign marginal cost shocks to the US ground coffee market. This was further developed in "Accounting for Incomplete Pass-Through" [6], with Dawit Zerom, which undertakes a structural econometric estimation of the sources of imperfect pass-through. The ground coffee industry provides a good laboratory for the study of pass-through of exogenous cost shocks for several reasons: green coffee beans represent at least 50 percent of costs of ground coffee manufacturers, green coffee beans are a fairly homogeneous input, and the world price of green coffee beans is subject to large weather shocks in the coffee-growing regions that can compellingly be treated as exogenous to US business cycle factors.

The paper presents a careful and skillful combination of data compilation and state-of-the-art econometrics, industrial organization theory, and computational methods. Emi and Dawit decompose imperfect pass-through into three potential sources: domestic cost components, desired markup adjustment, and nominal price adjustment costs. The paper finds that at the wholesale level, local costs reduce pass-through by 59 percent, mark-up adjustments reduce pass-through by 33 percent, and price adjustment costs have a negligible effect on pass-through after six quarters. Price adjustment costs, while of little importance in accounting for incompleteness of long-run pass-through, are found to be important in explaining the delayed pass-through in the short run. This paper, together with the Goldberg and Hellerstein (2013) study of the beer industry, represent the first attempts to incorporate price adjustment costs as a third determinant of incomplete cost pass-through in the context of structural estimation, and the results are consistent across the two studies.

The paper [6] also sheds some light on the different degree of cost pass-through at the wholesale and the retail levels. In particular, incomplete cost pass-through is shown to occur at the wholesale level; that is, changes in green coffee bean costs are incompletely passed through to wholesale ground coffee prices. By contrast, Emi and Dawit show that changes in wholesale ground coffee prices tend to be passed through to retail ground coffee prices fully and without much delay.

Emi also studies the differences between retail and wholesale pass-through in "Pass-Through in Retail and Wholesale" [3]. This time she works with a large panel dataset on weekly observations of prices for the year 2004 for a cross section of about 100 grocery items at the barcode level collected at 7,000 grocery stores operated by the largest supermarket chains in the United States. The dataset has close to 50 million price and quantity observations. She seems to have been the first to study this dataset in a macro context. An important aspect of Emi's dataset is that it has price observations for the same good at the same time at different grocery chains, whereas most of the related literature used data from a single grocery chain (the Dominick's Finer Foods data mentioned earlier).

Emi's focus in [3] is not on the extent to which cost shocks are passed through from the wholesale level to the retail level, but rather, what are the sources of retail price variations and are they related to shocks at the wholesale level? It turns out that only 16 percent of price changes are common across stores selling an identical item, which implies that only a small fraction of retail price variation is due to common cost shocks. Emi further finds that 65 percent of the price variation is common to stores within a particular retail chain, which suggests that the source of price fluctuations might be specific to shocks that the retail chain faces.

Emi's most recent contribution to the analysis of pass-through from costs to prices is "Informational Rigidities and the Stickiness of Temporary Sales" [17], with Eric Anderson, Benjamin Malin, Duncan Simester, and Jón Steinsson. This paper asks whether aggregate cost shocks are transmitted to retail prices via regular prices or sales. Although 95 percent of movements in prices are changes in sales prices, the paper provides evidence that aggregate cost shocks are mostly transmitted via changes in base prices. The empirical analysis is based on 195 weeks of scanner price data from 102 stores at a larger retailer that sells products in the grocery, health and beauty, and general merchandise categories. The central finding is that in a substantial fraction of cases, when the base wholesale price increases (that is, a cost shock for the retailer), the regular retail price responds quickly and completely while sales experience no reductions either in frequency or in size. On the contrary, discounts temporarily increase when regular retail prices increase, which the authors interpret as attempts to mask the associated regular price increase. The paper performs a number of robustness checks, including documenting that base retail prices respond more consistently than sales to changes in commodity price and to changes in unemployment, and by documenting that sales have a small contribution to overall inflation relative to base-price changes. This paper should change many views on the role of sales in the transmission of aggregate shocks.

Empirical Studies of Fiscal Stimulus

Establishing the size of the government spending multiplier is a fundamental question in macroeconomics, but despite a very large body of work, the answer remains controversial. Existing estimates on the fiscal multiplier are quite dispersed. Some studies suggest that the fiscal multiplier is close to zero, while others find that it is as large as two.

One difficulty in estimating the fiscal multiplier is to find truly exogenous changes in government spending. For example, military purchases are one plausible candidate for exogenous variations in government spending, but as Barro and Redlick (2011) note, there is likely to be insufficient variation in national-level US military spending in the last 50 years for a persuasive empirical test. Another problem with previous studies of the fiscal multiplier is that the output effects of government spending should depend on the nature of the monetary policy reaction. For example, if a study does not take into account how positive output effects can be reduced by the typical monetary response, the estimated size of the fiscal multiplier could be biased downward.

In "Fiscal Stimulus in a Monetary Union: Evidence from U.S. Regions" [13], Emi and Jón bring a fresh identification approach and new data to this long-standing debate. They sidestep the problem of insufficient national-level variation in military spending by showing that there has been sizable variation in regional military spending and those regional variations can thus be used to estimate the government spending multiplier. In addition, because the monetary policy reaction is common to all states, it is not a factor in explaining the differential effects on output across states. A further complication in estimating government spending multipliers is that their size depends on how government spending changes are financed. An advantage of Emi and Jón's empirical strategy in [13] is that regional military spending is financed by federal taxation and thus regions that receive a large chunk of military spending will not have associated tax payment structures that are different from regions that do not receive military spending.

For all of these reasons, considering variations in regional military spending and relating it to regional output variations should provide a more reliable estimate of the government spending multiplier than previous studies. In [13], Emi and Jón find that an increase in government spending equal to 1 percent of GDP increases output by 1.5 percent; that is, the government-spending multiplier measured in this way is 1.5.

However, this influential paper offers more than an instrument for measuring the "multiplier" effect of government purchases. As the authors point out, the multiplier for the effect of relatively higher purchases in one state on relative economic activity in that state need not be the same as the multiplier effect on national GDP of a nationwide increase in government purchases. The reason is that spillovers are likely to occur between states of the effects of increased purchases in any given state.

Emi and Jón [13] address the likely magnitude of the difference between the two multipliers by developing and analyzing a quantitative multi-region New

Keynesian general-equilibrium model. They use the paper to ask what the national multiplier would be in the case of a model parameterization that can account for their estimated relative state-level effects. The paper provides an excellent example of work that combines nonstructural empirical work with careful model-based analysis of what can be learned from the estimates.

The Effects of Monetary Policy

A key question in recent monetary policy debates is the extent to which central bank commitments about future policy, perhaps years into the future, can influence financial conditions and stimulate aggregate demand. The Federal Reserve and other central banks have been experimenting with "forward guidance" of this kind since the Great Recession.

Indeed, there is a "forward guidance puzzle" in which economic theory suggests that such guidance should be far more powerful than it actually seems to be. Specifically, some New Keynesian dynamic stochastic general equilibrium models find that a credible forward guidance commitment to maintain a fixed low nominal interest rate several years into the future will create a degree of output stimulus and/or inflation immediately that is difficult to regard as a realistic prediction. Of course, one possible resolution of the puzzle is that actual experience with forward guidance has not in fact involved credibly long-dated and such unconditional commitments, which is why actual forward guidance has had much more modest effects.

In their paper "The Power of Forward Guidance Revisited" [15], Emi and Jón, with Alisdair McKay, argue that this unrealistic implication of the simple New Keynesian models implying implausibly strong effects of forward guidance results from their assumption that each agent has a single intertemporal budget constraint. In turn, this assumption is the result of an underlying assumption (for modeling convenience) of complete financial markets and no borrowing constraints. They instead analyze the effects of a long-horizon commitment to a fixed nominal interest rate in a model that instead allows for the existence of uninsurable income risk and borrowing constraints. They find that while the effects of expectations about monetary policy at shorter horizons are similar to those predicted by the simpler model, the predicted effects of a long-lasting commitment to a fixed nominal interest rate are much weaker. Essentially, in the case of a household with a significant probability of facing a binding borrowing constraint over the next several quarters, expectations about monetary policy farther in the future do not affect its current ability to spend. In this way, the expectation of borrowing constraints substantially reduces the predicted effects on forward guidance—though it hardly implies that this policy tool is therefore irrelevant.

The paper [15] is important both as a contribution to a policy debate and as a methodological contribution on the use of New Keynesian models to assess alternative monetary policies. Its essential conclusion, that the effects of forward guidance are muted in more complex (and realistic) New Keynesian models, has

been supported by a number of subsequent analyses by other authors that consider generalizations of the basic model. The paper has also stimulated an active recent literature on "heterogeneous-agent New Keynesian models," which explores the implications for other aspects of macroeconomic dynamics of introducing income heterogeneity and borrowing constraints.

Another core question in macroeconomics is the effort to measure the effects of monetary policy shocks on the economy. One reason why answers to the question have remained controversial is because of the difficulty in distinguishing between exogenous changes in monetary policy and responses by the central bank to changes in economic conditions that have other sources. A recent strand of the literature looks at changes in financial market prices in a narrow time window around central bank policy announcements, based on the theory that these financial market movements can reveal how (or whether) the monetary policy announcement market movements indicate a change in the beliefs of market participants. In this approach, the size and direction of financial market movements can be taken as a measure of the monetary policy "shock" that has been revealed by the announcement. Regression of other variables on the time series of "shocks" identified in this way can then be taken to provide a measure of the causal effects of such shocks, as in Cook and Hahn (1989), Kuttner (2001), and Cochrane and Piazzesi (2002).

In "High Frequency Identification of Monetary Non-Neutrality: The Information Effect" [19], Emi and Jón note that this "high-frequency identification" strategy is subject to an important qualification, even if one grants that market movements during the short time window can only reflect information gleaned from the policy announcement. The issue is that new information from a central bank announcement might be of two types: a revelation that central bank policy will be different than would ordinarily be expected, given economic conditions; or alternatively, a revelation that the central bank's view of current economic conditions is different than the public expected. News of the former kind would correspond to a policy "shock." But to the extent that the central bank's unexpected view of the situation would be taken to reveal the central bank's superior information about economic conditions, such news should change people's own understanding of those conditions as well, and hence change the way they trade in financial markets for reasons unrelated to the implications for monetary policy.

Emi and Jón ask whether it is possible to separate "information effects" of monetary policy announcements of this latter sort from the effects of news about monetary policy. They study nominal interest rate changes observed in a 30-minute window around 106 scheduled Federal Reserve announcements between January 2000 and March 2014. In [19], they propose an estimate of the effects of monetary policy shocks taking into account the presence of information effects and to build and estimate a theoretical model that can explain the observed effects of Fed announcements.

The paper [19] first documents that Fed announcements shift short-term nominal and real rates almost one-for-one; that is, if the announcement results in a ten-basis-point increase in nominal short rates, then it also causes a ten-basis-point

increase in real short rates. This effect on real rates is observed not only for short-term rates, but also for longer term ones. Further, and also consistent with the related empirical literature, the paper documents that Fed announcements have little effect on expected inflation and that announcements that lead to an increase in nominal rates tend to be associated with increased expectations of future output growth. The latter empirical regularity is not easily reconciled with interpreting the news in Fed announcements as pure monetary policy shocks, since in canonical monetary models such shocks should lead to a downward revision of future output growth.

This interpretation problem motivates their development of a model in which Federal Reserve announcements can have both an information effect and a pure monetary policy shock, accompanied by estimation of the size of each component. Using the proposed model, they find that about two-thirds of the announcement shock represents news about future economic fundamentals and hence only one-third represents a pure monetary policy shock. They also find that, despite the great importance of the information effect, the observed responses to Fed announcements are consistent with a high degree of monetary non-neutrality in the US economy. These important results about fundamental questions in monetary economics are relevant not only for policy design but for understanding of the kinds of models that can best account for the nature of business fluctuations more generally.

Empirical Studies of Asset Pricing

One of the largest literatures for any question in economics is the search for an explanation for the equity premium puzzle, which refers to the large differential over time between the average return on US equities and the average return to short-term Treasury securities. In the early 1980s, Grossman and Shiller (1981) and Mehra and Prescott (1985) noted that the risk of holding stocks for a representative household should not be quantitatively significant because the covariance between consumption growth and equity returns in the post-World War II US economy was very low—in great part because the volatility of aggregate consumption growth is itself quite low. As a result, a standard asset pricing model would then imply that the compensation in return for holding equity rather than bonds should also be small.

Many theories have been proposed to explain the equity premium puzzle, and two recent candidate explanations have attracted particular attention. The first, most fully developed by Barro (2006), argues that the post-World War II sample underestimates the volatility of consumption because it does not include an example of the rare, large disasters that lead to large falls in consumption. The other, put forward by Bansal and Yaron (2004), argues that there are persistent shocks to both the long-run mean growth rate of consumption and the variance of its innovations, which short samples miss. Both stories are plausible, and large follow-up literatures have shown that, if their premises are true, they can explain the equity premium together with other related asset-pricing puzzles.

Barro and Ursúa (2010) have put together a remarkable dataset with annual consumption for 24 countries and more than 100 years. Emi and various coauthors have used these data to test the premises behind the two leading explanations of the equity premium.

In "Crises and Recoveries in an Empirical Model of Consumption Disasters" [11], Emi and Jón, together with Barro and Ursúa, assess the extent to which rare disasters can account for the equity premium. They make two main changes from previous work. First, previous work had assumed that disasters unfold quickly and lead to a permanent fall in consumption. However, here the authors find that in the disasters in their sample, the trough occurred only six years after the disaster hit, and that more than half of the initial reduction in consumption was eventually reversed. Second, the authors assume Epstein-Zin preferences with an intertemporal elasticity of substitution of two, which suggests that people have a strong preference for early resolution of uncertainty, and so are very averse to extended disasters and their uncertain recoveries. Given these assumptions, along with a modest degree of risk aversion, the rare disasters of the type that have been historically observed would suffice to explain the equity premium.

In another paper on this topic, "Growth-Rate and Uncertainty Shocks in Consumption: Cross-Country Evidence" [16], Emi and Jón, with Dmitriy Sergeyev, use a subset of the same long panel of data on consumption to reassess the long-run risks model of Bansal and Yaron (2004). They allow for shocks to both the country-specific growth rate of consumption and also a world growth factor and their respective variances. They find that filtered estimates of the world growth rate track many of the medium-term fluctuations in macroeconomic variables that have been identified so far: the post-World War II productivity speed up, the slowdown in productivity after 1970, the Great Moderation from the 1980s to the 2000s, and the more recent increase in volatility. Again using Epstein-Zin preferences, Emi and her coauthors show that with a coefficient of intertemporal elasticity of 1.5 and a risk aversion coefficient of 6.5, the model can fit the average equity premium. These results provide validation for the long-run risks model, which had previously been untested in its key premise.

These papers illustrate Emi's ability to go after big questions in the literature, to perceive testable implications in the theory, to bring different data to bear than had been previously used, and ultimately to provide more convincing answers. It is unlikely that these papers will be the last word on this important and controversial topic; for example, the methods used to filter world and country-specific growth rates can be sensitive to underlying assumptions. Still, the methodology in these papers represents an important advance over previous work.

Conclusion

The examples of Emi's work described above are not exhaustive, but should suffice to illustrate some of her characteristic concerns. All of her work has been

driven not simply by a belief that careful measurement matters, but by close attention to subtle issues regarding the inferences that can legitimately be drawn from the available measurements.

Emi and Jón's views about the appropriate methodology for empirical work are most clearly enunciated in their paper on "Identification in Macroeconomics" [18] in the Summer 2018 issue of this journal. Here, they discuss why it has been so difficult to settle questions about the effects of monetary and fiscal policies and also to stress the limitations of two seemingly straightforward approaches. "Direct causal inference" seeks to find examples of exogenous changes in policy in the historical record and measure what happened. But as Emi and Jón point out, truly exogenous policy changes are relatively scarce, and those that can be observed seldom involve the kind of change that is relevant for policy development, raising questions of external validity from the available "natural experiments."

Accordingly, an influential alternative approach argues that one can only hope to answer questions about counterfactual policies using a fully specified structural model of the macro-economy. Many researchers in the real business cycle tradition further propose that the quantitative realism of such models should be validated by comparing the predicted values of various unconditional moments (the overall variability of aggregate investment spending relative to the overall variability of real GDP, and so on) to the empirical values of these moments. The advantage of a focus on matching the values of unconditional moments is that these quantitative targets can be defined in a way that is independent of any particular theoretical structure. But as Emi and Jón note, this approach has the disadvantage that predictions for the statistics in question depend on the simultaneous specification of a large number of aspects of a macro model. One can only judge the model as a complete whole to be successful or unsuccessful in matching reality.

Emi and Jón argue instead for the desirability of focusing on the measurement of what they call "identified moments," by which they mean estimates of the effects of particular types of identified disturbances. This approach differs from "direct causal inference" insofar as it admits that the responses that can be measured will not generally provide a direct answer to the questions about counterfactual policies that one actually wishes to answer; instead, one measures responses to disturbances that can be identified using assumptions that are as credible as possible and then uses the answers to these questions to discipline the parameterization of the structural models that will be used to answer the questions of real interest. At the same time, this approach differs from unconditional matching of statistical moments in that the quantitative targets that the structural model is required to match are selected so that they allow diagnosis of the quantitative realism of certain parts of the model, rather than depending equally on all aspects of the model specification. The paper provides a powerful case for the fruitfulness of this alternative approach, and shows how it has guided Emi and Jón's own work on the effects of monetary and fiscal policy, discussed above.

Emi's methodological approach to research is a signature contribution, as the many examples discussed in this paper should help to convey. She has demonstrated

that macro models have rich implications for the underlying dynamics of the economy. She has focused on testing these fine aspects of the empirical record as the most reliable way of determining which models best describe the world and hence can best be relied upon as guides to policy. She has shown extraordinary ingenuity in connecting micro data to macro models and has taken painstaking care in developing new data when needed. But despite Emi's frequent emphasis on the importance of careful scrutiny of fine-grained data, her work never loses sight of the big questions about the nature of economic fluctuations and the effects of policy that macroeconomic models are intended to answer. This combination of care in precisely defining what one really can measure while marshaling all possible evidence to answer questions of first-order importance is what has made her work so highly influential.

Some of the qualities that have made Emi's work influential arose early and naturally from her curiosity about metrics and her commitment to measurement. But these qualities could have been just as easily applied to small questions. Emi's research has been *transformative* because it has demonstrated that these qualities are also applicable to big, "messy" questions in macroeconomics, where the available data often seemed to be limited, and before her work, it was not obvious how to address these questions with more granular data. Emi's work shows how to reach the big questions, building from models and data that look at them "up close" so that we can see them clearly.

References

Athey, Susan, Kyle Bagwell, and Chris Sanchirico. 2004. "Collusion and Price Rigidity." *Review of Economic Studies* 71 (2): 317–49.

Bansal, Ravi, and Amir Yaron. 2004. "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles." *Journal of Finance* 59 (4): 1481–1509.

Barro, Robert J. 2006. "Rare Disasters and Asset Markets in the Twentieth Century." *Quarterly Journal of Economics* 121 (3): 823–66.

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

Barro, Robert, and José Ursúa. 2010. "Barro-Ursua Macroeconomic Data." https://scholar.harvard.edu/barro/publications/barro-ursua-macroeconomic-data.

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

Chari, V. V., and Patrick J. Kehoe. 1990. "Sustainable Plans." Journal of Political Economy 98 (4): 783–802.

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

Cook, Timothy, and Thomas Hahn. 1989. "The Effect of Changes in the Federal Funds Rate Target on Market Interest Rates in the 1970s." *Journal of Monetary Economics* 24 (3): 331–51.

Goldberg, Pinelopi Koujianou, and Rebecca Hellerstein. 2013. "A Structural Approach to Identifying the Sources of Local Currency Price Stability." *Review of Economic Studies* 80 (1): 175–210.

Golosov, Mikhail, and Robert E. Lucas Jr. 2007. "Menu Costs and Phillips Curves." *Journal of Political Economy* 115 (2): 171–99.

Grossman, Sanford J., and Robert J. Shiller. 1981. "The Determinants of the Variability of Stock Market Prices." *American Economic Review* 71 (2): 222–27.

Kuttner, Kenneth N. 2001. "Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market." *Journal of Monetary Economics* 47 (3): 523–44.

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

Ng, Serena. 2015. "An Interview with Emi Nakamura." *CSWEP News*. https://www.aeaweb.org/content/file?id=521.

Rampell, Catherine. 2013. "Outsource Your Way to Success." *New York Times Magazine*, November 5. https://www.nytimes.com/2013/11/10/magazine/outsource-your-way-to-success.html.

Ravn, Morten, Stephanie Schmitt-Grohé, and Martín Uribe. 2006. "Deep Habits." *Review of Economic Studies* 73 (1): 195–218.

Watts, Harold W. 1991. "Distinguished Fellow: An Appreciation of Guy Orcutt." *Journal of Economic Perspectives* 5 (1): 171–79.

Journal of Economic Perspectives—Volume 34, Number 1—Winter 2020—Pages 240–247

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., St. Paul, MN 55105.

Smorgasbord

Bryce Pardo, Jirka Taylor, Jonathan P. Caulkins, Beau Kilmer, Peter Reuter, and Bradley D. Stein have co-authored an e-book on *The Future of Fentanyl and Other Synthetic Opioids* (RAND Institute 2019, https://www.rand.org/pubs/research_reports/RR3117.html). "Although the media and the public describe an opioid epidemic, it is more accurate to think of it as a series of overlapping and interrelated epidemics of pharmacologically similar substances—the opioid class of drugs. … The first wave was prescription opioids, the second wave was heroin, and the third—and ongoing—wave is synthetic opioids, such as fentanyl. … Most of the fentanyl and novel synthetic opioids in U.S. street markets—as well as their precursor chemicals—originate in China, where the regulatory system does not effectively police

■ *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.34.1.240.

the country's expansive pharmaceutical and chemical industries. According to federal law enforcement, synthetic opioids arrive in U.S. markets directly from Chinese manufacturers via the post, private couriers (e.g., UPS, FedEx), cargo, by smugglers from Mexico, or by smugglers from Canada after being pressed into counterfeit prescription pills. …Recent RAND Corporation research identified multiple Chinese firms that are willing to ship 1 kg of nearly pure fentanyl to the United States for $2,000 to $5,000. In terms of the morphine-equivalent dose (MED; a common method of comparing the strength of different opioids), a 95-percent pure kg of fentanyl at $5,000 would generally equate to less than $100 per MED kg. For comparison, a 50-percent pure kg of Mexican heroin that costs $25,000 when exported to the United States would equate to at least $10,000 per MED kg. Thus, heroin appears to be at least 100 times more expensive than fentanyl in terms of MED at the import level. … For reference, if the total U.S. heroin market was on the order of 45 pure metric tons. … before fentanyl and if fentanyl is 25 times more potent than heroin, then it would only take 1,800 1-kg parcels to supply the same amount of MEDs to meet the demand for the entire U.S. heroin market."

The *World Development Report 2020* is subtitled "Trading for Development in the Age of Global Value Chains" (World Bank, October 2019, https://www.worldbank.org/en/publication/wdr2020). "International trade expanded rapidly after 1990, powered by the rise of global value chains (GVCs). This expansion enabled an unprecedented convergence: poor countries grew faster and began to catch up with richer countries. Poverty fell sharply. These gains were driven by the fragmentation of production across countries and the growth of connections between firms. Parts and components began crisscrossing the globe as firms looked for efficiencies wherever they could find them. Productivity and incomes rose in countries that became integral to GVCs—Bangladesh, China, and Vietnam, among others. The steepest declines in poverty occurred in precisely those countries. Today, however, it can no longer be taken for granted that trade will remain a force for prosperity. Since the global financial crisis of 2008, the growth of trade has been sluggish, and the expansion of GVCs has slowed. …At the same time, two potentially serious threats have emerged to the successful model of labor-intensive, trade-led growth. First, the arrival of labor-saving technologies such as automation and 3D printing could draw production closer to the consumer and reduce the demand for labor at home and abroad. Second, trade conflict among large countries could lead to a retrenchment or a segmentation of GVCs."

The *World Trade Report* focuses on "the future of services trade" (World Trade Organization, 2019, https://www.wto.org/english/res_e/publications_e/wtr19_e.htm). "While the value of goods exports has increased at a modest 1 per cent annually since 2011, the value of commercial services exports has expanded at three times that rate, 3 per cent. The services share of world trade has grown from just 9 per cent in 1970 to over 20 per cent today—and this report forecasts that services could account for up to one-third of world trade by 2040. This would represent a 50 per cent increase in the share of services in global trade in just two decades. There is a common perception that globalization is slowing down. But if the growing wave

of services trade is factored in—and not just the modest increases in merchandise trade—then globalization may be poised to speed up again."

Venkatraman Anantha Nageswaran and Gulzar Natarajan explore "India's Quest for Jobs: A Policy Agenda" (Carnegie India, September 2019, https://carnegieindia.org/2019/10/03/india-s-quest-for-jobs-policy-agenda-pub-79967). "By 2020, India is expected to be the youngest country in the world, with a median age of twenty-nine, compared to thirty-seven for the most populous country, China.…The burgeoning youth population has led to an estimated 10–12 million people entering the workforce each year. In addition, the rapidly growing economy is transitioning away from the agricultural sector, with many workers moving into secondary and tertiary sectors. Employing this massive supply of labor is, perhaps, the biggest challenge facing India.…India is often considered one of the most difficult places to start and run a business.…One of the biggest hurdles that potential enterprises in India face is the complexity of the registration system—all enterprises must register separately with multiple entities of the state and central governments.… Further, there are registrations specific to sector or occupational categories—for example, manufacturing enterprises with more than ten employees must register with the labor department under the Factories Act.…According to current labor laws, service enterprises and factories must maintain twenty-five and forty-five registers, respectively, and file semi-annual and annual returns in duplicate and in hard copy. Furthermore, regular paperwork tends to be convoluted; salary and attendance documents should be simple but instead require tens of entries.…All these requirements add up to impose prohibitive costs that reduce the success of these businesses." This paper can be read as a complement to the three-paper "Symposium on India" in this issue.

Adel Abdellatif, Paola Pagliani, and Ellen Hsu discuss "Leaving No One Behind: Towards Inclusive Citizenship in Arab Countries" (July 2019, Arab Human Development Report Research Paper, http://www.arab-hdr.org/UNDP_Citizenship_and_SDGs_report_web.pdf). "Unaccountable and unresponsive public institutions as well as perceived widespread corruption often drive exclusion and disenfranchisement for large segments of the population.…Trust in elected bodies, those that should be in charge of redesigning the social contract, is particularly low. Lack of trust is also reflected in low electoral turnouts—below 50 percent in most countries.…Perceptions of ineffective institutions seem confirmed by stagnating or narrowly based economic structures, high unemployment, young people facing difficult prospects to secure their future and uneven provision of social services and social protection nets. Unemployment, averaging 10 percent, almost double the world average, disproportionately affects young people, at 25 percent.…84% of the population is affected by or at risk of water scarcity. The decline of arable land and the dependency on food imports expose the population to risks of food insecurity…"

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019 was awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer "for their experimental approach to alleviating global poverty." As background,

the Nobel committee published "Scientific Background: Understanding Development and Poverty Alleviation" (October 14, 2019, https://www.nobelprize.org/uploads/2019/10/advanced-economicsciencesprize2019.pdf). "This year's Prize in Economic Sciences rewards the experimental approach that has transformed development economics. …First, in the mid-1990s, Kremer and his co-authors launched a set of randomized controlled trials on schooling in Kenya. In effect, their approach amounted to splitting up the question of how to boost human capital in low-income countries into smaller and more manageable specific topics, each of which could be rigorously studied via a carefully designed field experiment. Soon thereafter, Banerjee and Duflo, often together with Kremer or others, broadened the set of educational topics and expanded the scope of the research to other areas, including health, credit and agriculture. Second, in a series of contributions, Banerjee and Duflo articulated how pieces from such microeconomic studies can help us get closer to solving the broad development puzzle: what explains the enormous difference in per-capita income across countries?…A deeper understanding of the development problem thus requires an explanation of why some firms and individuals do not take advantage of the best available opportunities and technologies. Banerjee and Duflo further argued that these misallocations can be traced back to various market imperfections and government failures. …Finally, by designing new experimental research methods and by addressing the key challenge of generalizing results from a specific experiment—i.e., the issue of external validity—the Laureates firmly established this transformed approach to development economics."

Universal Basic Income

Melissa S. Kearney and Magne Mogstad have written "Universal Basic Income (UBI) as a Policy Response to Current Challenges" (Aspen Institute Economic Strategy Group, August 23, 2019, https://www.brookings.edu/wp-content/uploads/2019/08/UBI-ESG-Memo-082319.pdf). "First, some view a UBI [universal basic income] as a reasonable response to growing inequality, to stem both economic and political unease. …Second, some worry about the widespread elimination of well-paying jobs for many workers in the U.S. due to robots and other technological advancements. For this reason, the idea seems to have caught on among a number of tech futurist personalities. …A third, very distinct motivation for a UBI scheme is to streamline the current complicated and sometimes counterproductive system of U.S. transfer programs. …We view a UBI to be a sub-optimal, and possibly harmful, policy response to all three of these challenges. A UBI in its most basic form would be massively expensive yet do little to reduce inequality or advance opportunity. Devoting that level of spending to targeted benefits, focusing on the poorest and those hardest hit by ongoing economic forces, and polices dedicated to human capital development instead of mere redistribution would produce a much greater social return than a UBI."

The *Annual Review of Economics* includes a three-paper "Symposium: Universal Basic Income" (August 2019, https://www.annualreviews.org/toc/economics/11/1): "Universal Basic Income: Some Theoretical Aspects," by Maitreesh Ghatak and François Maniquet; "Universal Basic Income in the United States and Advanced Countries," by Hilary Hoynes and Jesse Rothstein; and "Universal Basic Income in the Developing World," by Abhijit Banerjee, Paul Niehaus, and Tavneet Suri. For example, Banerjee, Niehaus, and Suri write: "A central question about UBI is whether universality is in fact efficient. For any given budget, is it better to spread those resources evenly or to give larger amounts to the poorest?…We suspect that universality has several under-appreciated benefits, and targeting several under-appreciated limitations.…Government capacity to implement nuanced targeting schemes is often limited, particularly so in the poorest areas where it is most important to get it right. In cases like these, making eligibility universal may have a modest effect on the realized incidence of benefits while at the same time substantially reducing the scope for corruption and other abuses of power."

Symposia

Dædalus has published a 12-paper symposium (plus an introduction) about "Improving Teaching: Strengthening the College Learning Experience," edited by Sandy Baum and Michael McPherson (Fall 2019, https://www.amacad.org/sites/default/files/daedalus/downloads/Daedalus_Fa2019_Book.pdf). They write in their introductory essay: "An odd feature of the public policy discussion of higher education is the near absence of attention to the quality of teaching.…Instead, questions about college admissions, pricing and cost, debt, and financial returns dominate the news and policy discussion. These are worthy topics of study, but they sidestep examination of what goes on inside the `black box' of teaching and learning that college students actually experience.…An observer from another planet visiting American Ph.D. programs might well conclude that the graduate students there are being prepared for full-time careers in academic research.…Yet after graduating, typical faculty members in the United States actually spend the majority of their professional time on undergraduate teaching and related activities, spending less than one-quarter of their time on graduate instruction and research combined. The "theory" that would justify this mismatch between what faculty are prepared for and what they actually do is that the hard part of being a good teacher is knowing the subject matter, and the rest can be picked up 'on the job.' This is not an assumption we would readily accept in other professions like aviation or surgery…"

The *Russell Sage Foundation Journal of the Social Sciences* has published a 10-paper symposium, edited by Erica L. Groshen and Harry J. Holzer, on the general theme of "Improving Employment and Earnings in Twenty-First Century Labor Markets" (December 2019, https://www.rsfjournal.org/content/5/5). For a sense of the contents, here are titles and authors for the first five papers: "From Immigrants to

Robots: The Changing Locus of Substitutes for Workers," by George J. Borjas and Richard B. Freeman; "Public Universities: The Supply Side of Building a Skilled Workforce," by John Bound, Breno Braga, Gaurav Khanna, and Sarah Turner; "Wages and Hours Laws: What Do We Know? What Can Be Done?" by Charles C. Brown and Daniel S. Hamermesh; "Unions, Worker Voice, and Management Practices: Implications for a High-Productivity, High-Wage Economy," by Thomas A. Kochan and William T. Kimball; "Making Ends Meet: The Role of Informal Work in Supplementing Americans' Income," by Katharine G. Abraham and Susan N. Houseman.

Conversations with Economists

Tyler Cowen conducts one of his "Conversations with Tyler" with Hal Varian in "Hal Varian on Taking the Academic Approach to Business" (Medium.com, June 19, 2019, https://medium.com/conversations-with-tyler/tyler-cowen-hal-varian-google-9326e0d59ba2). TC: "How will 5G change my world?" HV: "Basically, you should think of 5G as Wi-Fi everywhere so that you've got a high-speed communication without having to go through any sort of special operations. …When you look at technologies like autonomous vehicles and things like that, they're dealing with vast amounts of information. It's often stored and manipulated locally, but sometimes it needs to be shared. Doing that kind of sharing will be easier if you have high-bandwidth 5G technology. But realistically speaking, for most of what you're going to be doing, it will just save you a small amount of time." TC: "Why are textbooks still priced so high?" HV: "They are priced remarkably high, and it's a situation where I really would like to see lower prices because, obviously, there's a durable goods monopoly problem there. As you have more and more competition from previous editions, each of the new editions has to differ markedly from the old edition to support the pricing model. But that's getting harder and harder to do. In fact, a friend of mine once told me, 'Having a successful textbook is like being married to a very wealthy person you don't like much anymore.'"

Catherine L. Kling and Fran Sussman share "A Conversation with Maureen Cropper" in the *Annual Review of Resource Economics* (October 2019, 11, pp. 1–18, https://www.annualreviews.org/doi/pdf/10.1146/annurev-resource-100518-093858). For example, I had not known that Cropper started as a monetary economist. "Frankly, my interests at the time were really in monetary economics, so I took several courses at the Cornell Business School, including courses in portfolio theory. My dissertation was on bank portfolio selection with stochastic deposit flows. …At this time, I was not doing anything in environmental economics. In fact, my first job offer was from the NYU Business School. The reason I went into environmental economics is that I met Russ Porter in graduate school. …We decided that we would go on the job market together and looked for a place that would hire two economists. We wound up at the University of California, Riverside, which at the time was the birthplace of the *Journal of*

Environmental Economics and Management (*JEEM*). …It was going to UC Riverside that really caused me to switch fields and go into environmental economics. … There are moments when I wonder what would've happened if I had gone to the NYU Business School instead of UC Riverside. There are many situations when it is, to some extent, a matter of chance how things will unfold. Would I do anything differently? No, I don't think so, not really."

David A. Price interviews Emmanuel Farhi (*Econ Focus*, Regional Federal Reserve Bank of Richmond, Second/Third Quarter 2019, pp. 18–23, https://www.richmondfed.org/-/media/richmondfedorg/publications/research/econ_focus/2019/q2-3/interview.pdf). "If you look at the world today, it's very much still dollar-centric. …The U.S. is really sort of the world banker. As such, it enjoys an exorbitant privilege and it also bears exorbitant duties. Directly or indirectly, it's the pre-eminent supplier of safe and liquid assets to the rest of the world. It's the issuer of the dominant currency of trade invoicing. And it's also the strongest force in global monetary policy as well as the main lender of last resort. If you think about it, these attributes reinforce each other. The dollar's dominance in trade invoicing makes it more attractive to borrow in dollars, which in turn makes it more desirable to price in dollars. And the U.S. role as a lender of last resort makes it safer to borrow in dollars. That, in turn, increases the responsibility of the U.S. in times of crisis. All these factors consolidate the special position of the U.S. But I don't think that it's a very sustainable situation. More and more, this hegemonic or central position is becoming too much for the U.S. to bear. …In my view, there's a growing and seemingly insatiable global demand for safe assets. And there is a limited ability to supply them. In fact, the U.S. is the main supplier of safe assets to the rest of the world. As the size of the U.S. economy keeps shrinking as a share of the world economy, so does its ability to keep up with the growing global demand for safe assets. The result is a growing global safe asset shortage. It is responsible for the very low levels of interest rates that we see throughout the globe. And it is a structural destabilizing force for the world economy. …Basically, I think that the role of the hegemon is becoming too heavy for the U.S. to bear. And it's only a matter of time before powers like China and the eurozone start challenging the global status of the dollar as the world's pre-eminent reserve and invoicing currency. It hasn't happened yet. But you have to take the long view here and think about the next decades, not the next five years. I think that it will happen."

Discussion Starters

William H. Shrank, Teresa L. Rogstad, and Natasha Parekh discuss "Waste in the US Health Care System: Estimated Costs and Potential for Savings" (*Journal of the American Medical Association*, October 7, 2019, https://jamanetwork.com/journals/jama/fullarticle/2752664). "In this review based on 6 previously identified domains of health care waste, the estimated cost of waste in the US health

care system ranged from $760 billion to $935 billion, accounting for approximately 25% of total health care spending.…Computations yielded the following estimated ranges of total annual cost of waste: failure of care delivery, $102.4 billion to $165.7 billion; failure of care coordination, $27.2 billion to $78.2 billion; overtreatment or low-value care, $75.7 billion to $101.2 billion; pricing failure, $230.7 billion to $240.5 billion; fraud and abuse, $58.5 billion to $83.9 billion; and administrative complexity, $265.6 billion."

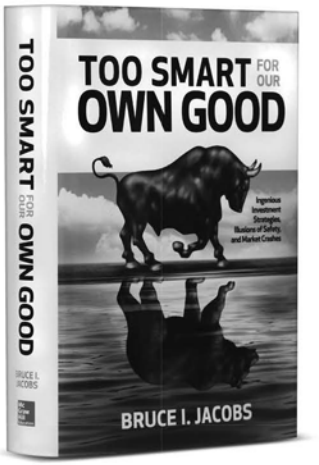
The Health Effects Institute has authored the *State of Global Air 2019* (https://www.stateofglobalair.org/sites/default/files/soga_2019_report.pdf): "Air pollution (ambient PM2.5, household, and ozone) is estimated to have contributed to about 4.9 million deaths (8.7% of all deaths globally) and 147 million years of healthy life lost (5.9% of all DALYs [disability-adjusted life years] globally) in 2017. The 10 countries with the highest mortality burden attributable to air pollution in 2017 were China (1.2 million), India (1.2 million), Pakistan (128,000), Indonesia (124,000), Bangladesh (123,000), Nigeria (114,000), the United States (108,000), Russia (99,000), Brazil (66,000), and the Philippines (64,000).…Air pollution collectively reduced life expectancy by 1 year and 8 months on average worldwide, a global impact rivaling that of smoking. This means a child born today will die 20 months sooner, on average, than would be expected in the absence of air pollution."

Jason D. Delisle and Preston Cooper offer a short essay on "International Higher Education Rankings: Why No Country's Higher Education System Can Be the Best" (American Enterprise Institute, August 2019, https://www.aei.org/research-products/report/higher-education-rankings-no-countrys-system-best/). "In England, where the vast majority of the country's population is concentrated, universities charge undergraduate students tuition of up to $11,856, making English universities some of the most expensive in the world. That is why the United Kingdom ranks last on subsidies in our analysis, with just 26 percent of higher education funding derived from public sources. However, Britain's student loan program complicates this high-tuition, low-subsidy story. To enable students to afford these high fees, the government offers student loans that fully cover tuition. Ninety-five percent of eligible students borrow. Repayment is income contingent; new students pay back 9 percent of their income above a threshold for up to 30 years, after which remaining balances are forgiven. Despite the lengthy term, the program is heavily subsidized: The government estimates that just 45 percent of borrowers who take out loans after 2016 will repay them in full.…England's high-resource, high-tuition model is relatively new. Until 1998, English universities were tuition-free, with the government directly appropriating the vast majority of higher education funding.…In 1998, the center-left government of Tony Blair began allowing institutions to charge tuition to supplement their direct government funding. At the same time, the government expanded its student loan program and introduced income-contingent repayment. Over the next two decades, university enrollments and funding both surged, and today the United Kingdom ranks among the top nations for both resources and attainment."

Too Smart for Our Own Good

Ingenious Investment Strategies, Illusions of Safety, and Market Crashes

Bruce I. Jacobs



One of today's leading financial thinkers, Bruce I. Jacobs, examines recent financial crises—including the 1987 stock market crash, the 1998 collapse of the hedge fund Long-Term Capital Management, the 2007–2008 credit crisis, and the European debt crisis—and reveals the common threads that explain these market disruptions. In each case, investors in search of safety were drawn to novel strategies that were intended to reduce risk but actually magnified it—and blew up markets. Until we manage risk in responsible ways, major crises will always be just around the bend. *Too Smart for Our Own Good* is a big step toward smarter investing—and a better financial future for everyone.

"Bruce Jacobs explains when a crash is likely: It's when the economy is strong and risks appear to be low. Buy this book today and be forewarned."
—Elroy Dimson, Professor of Finance, Cambridge Business School

"Bruce Jacobs's insightful analyses of financial crises will alert readers to how some financial instruments and strategies can mask investment risk and lead to excessive leverage. Investors and financial institutions would do well to heed the warnings in this book."
—Frank J. Fabozzi, Professor of Finance, EDHEC Business School, and Editor, *The Journal of Portfolio Management*

"Bruce Jacobs takes a close look at financial blowups over four decades and finds a common element: risk management and investment strategies that appear benign at the micro level but pose dire systemic risks at the macro level."
—Greg Feldberg, Director of Research, US Financial Crisis Inquiry Commission

"*Too Smart for Our Own Good* is a remarkable combination of decades of hands-on wisdom from a great investor with astute analytical insight born of detailed research—on a topic that is vital not only to the world of finance, but also to the world at large."
—Geoffrey Garrett, Dean, The Wharton School

"The increasing frequency of market crashes is a clarion call for a thorough investigation of the causes of market fragility. *Too Smart for Our Own Good* offers a critical analysis that is of paramount importance for all of us."
—Michael Gibbons, Deputy Dean, and Professor of Finance, The Wharton School

About the Author

Bruce I. Jacobs is co-founder, co-chief investment officer, and co-director of research at Jacobs Levy Equity Management. He is co-author, with Ken Levy, of *Equity Management: The Art and Science of Modern Quantitative Investing*. Jacobs serves on the Advisory Boards of the *Journal of Portfolio Management* and *Journal of Financial Data Science*, and has served on the *Financial Analysts Journal* Advisory Council. He holds a Ph.D. in finance from The Wharton School.

Mc Graw Hill Education
Visit: www.mhprofessional.com
ISBN: 9781260440546
Available in print, ebook, and audiobook formats

THE COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION

CSWEP

Advancing the Status of Women in the Economics Profession

- Publishes an annual survey of representation of women in economics
- Offers mentoring workshops for junior faculty
- Conducts programs at the AEA and regional meetings
- Co-sponsors the summer economics fellows program
- Publishes a newsletter with professional development advice

Sign up to receive our newsletter
info@cswep.org

www.CSWEP.org

@AEACSWEP

Committee on the Status of Minority Groups in the Economics Profession

AMERICAN ECONOMIC ASSOCIATION

• Established 1968 •

SUPPORTING DIVERSITY IN ECONOMICS

The Committee on the Status of Minority Groups in the Economics Profession (CSMGEP) was established by the American Economic Association (AEA) in 1968 to increase the representation of minorities in the economics profession, primarily by broadening opportunities for the training of underrepresented minorities.

CSMGEP Programs

- Summer Economics Fellows Program
- Mentoring Program
- Summer Training Program

www.csmgep.org

DON'T MISS...

CTREE 2020

The Tenth Annual Conference on Teaching and Research in Economic Education

Plenary Speakers Include:

Lisa Cook
Michigan State University

Dean Karlan
Northwestern University

Betsey Stevenson
University of Michigan

Speaker TBA
Federal Reserve Board

May 27–29, 2020
Chicago, Illinois
Westin Michigan Avenue
www.aeaweb.org/ctree/2020

CTREE is hosted by the AEA Committee on Economic Education in conjunction with the *Journal of Economic Education* and the Federal Reserve Bank of Chicago.

JOE NETWORK

Designed for Economists ... by Economists

The *Preferred* Hiring Tool for the Economics Job Market

1,700+
POSITIONS FILLED

5,100+
CANDIDATES FROM AROUND THE WORLD

150k+
REFERENCE LETTER REQUESTS

No other placement program offers a more comprehensive way to match high-caliber candidates with sought-after economics positions.

AEA's JOE Network automates all hiring tasks including unique management of the faculty reference letter-writing process. All users can share materials, communicate confidentially, and easily manage their files and data within a dedicated and secure area of the AEA website.

AMERICAN ECONOMIC ASSOCIATION

www.aeaweb.org/JOE

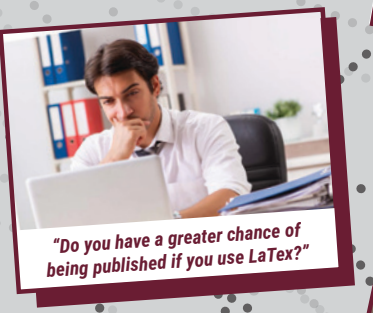
ASK • ADVISE • SHARE

Join the conversation!

New features include:
- *Multiple-mode posting option*
- *Polls • Inline images*

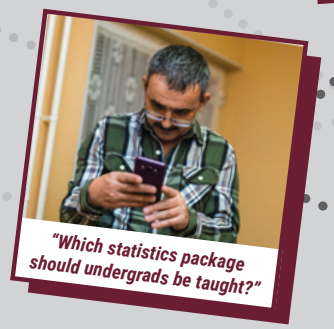
"Have low interest rates led to excessive risk taking?"

"What are the best sources for job market advice?"



"Do you have a greater chance of being published if you use LaTex?"

"What are some dissertation-worthy topics in market design?"



"Which statistics package should undergrads be taught?"

www.aeaweb.org/econspark

Participate in a community where economists at all career levels learn from and advise each other in a professional and supportive online environment.

AMERICAN ECONOMIC ASSOCIATION

ECON spark

AEA Discussion Forum

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 Journal of

Economic Perspectives

Winter 2020, Volume 34, Number 1

Symposia

Economics of India

Assimilation of Refugees

Electricity in Developing Countries

Articles

Features

AMERICAN ECONOMIC ASSOCIATION