The Journal of

Economic Perspectives

A journal of the
American Economic Association

Spring 2023

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), Spring 2023, Vol. 37, No. 2. The JEP is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. 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 USA.

The Journal of
Economic Perspectives

Contents

Volume 37 • Number 2 • Spring 2023

Symposia

Spatial and Urban Economics

Universal Health Insurance

Economics of Mental Health

Article

Features

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

Journal of Economic Perspectives

Advisory Board

Journal of Economic Perspectives—Volume 37, Number 2—Spring 2023—Pages 3–28

Economic Activity across Space: A Supply and Demand Approach

Treb Allen and Costas Arkolakis

The spatial distribution of people is incredibly concentrated: 8 percent of the US population lives in the ten largest US cities, but those cities take up less than 0.1 percent of total US land area. Why this concentration? More generally, what determines the distribution of people and economic activity across space? And how can economic policies affect the spatial distribution of economic activity? This essay will show that these questions can be answered through the familiar lens of supply and demand curves.

We begin by applying this intuition to the well-known Rosen-Roback framework (Rosen 1979; Roback 1982). But as we will discuss, the distribution of economic activity in this early spatial model depends only on local geography, not on what happens to other regions. For example, a change in one location—say, a large infrastructure investment that improves its productivity—is predicted to have an identical impact on all other locations, regardless of where they are. Thus, intuitive spatial features like where a location is located on a map and who its neighbors are entirely absent: it is a spatial model where space does not matter.

In reality, spatial linkages create rich interactions between locations. One implication of these interactions is that a large infrastructure investment that improves the productivity in one location will have greater impacts on close-by locations than locations further away. To account for such spatial linkages, we extend the intuition of the Rosen-Roback model to modern economic geography frameworks

■ *Treb Allen is Distinguished Associate Professor of Economics and Globalization, Dartmouth College, Hanover, New Hampshire. Costas Arkolakis is Professor of Economics, Yale University, New Haven, Connecticut. Their email addresses are treb.allen@dartmouth.edu and costas.arkolakis@yale.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.3.

where locations are connected through the flow of goods, based on our earlier work in Allen and Arkolakis (2014). In this framework, the economic fate of a location depends not only on its own "local" geography but also on the local geography of its neighbors, the effect of which is mediated by the strength of the economic ties, creating a "global geography." Despite this added complexity, we show the same tools based on supply and demand used to understand predictions of the earlier Rosen-Roback framework extend readily to a globally integrated world.

This globally integrated framework can be applied to understand both the direct and indirect impacts of real world economic policies that change either the local or global geography. We discuss how the framework can be applied to spatial data, while also highlighting the most common pitfalls and offering strategies for traversing them. Finally, we provide a brief overview of the many ways in which this framework has been applied thus to understanding the spatial distribution of economic activity, as well as pointing out several interesting and still unexplored questions for future researchers. To keep the discussion as straightforward and accessible as possible, we relegate all mathematical details and derivations to the Appendix, where we also provide a companion Matlab toolkit to help researchers apply these techniques on their own.

Understanding the Spatial Distribution of Economic Activity through the Lens of Supply and Demand

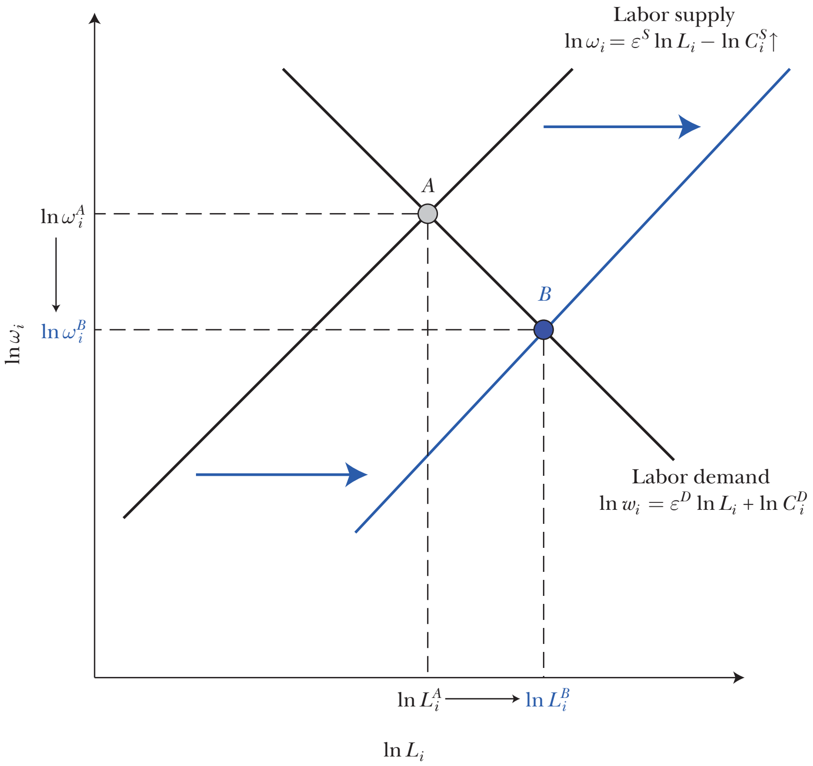
We now discuss the Rosen-Roback framework. Consider a world comprising many different locations. These locations each have their own "local" geography. The "local" geography of a location includes a whole host of things, from natural, geographic features like the climate, elevation, and natural beauty, to other less tangible characteristics of a location like the quality of its political institutions. Local geography can affect the spatial distribution of economic activity in two ways. First, it can affect the desire of people to live in a location and hence labor supply (we will call such factors "amenities"). Second, it can affect how productive people are in a location and hence labor demand (we will call such factors "productivities").

Figure 1 illustrates the spatial equilibrium that results for labor demand and labor supply in this market. The labor market envisaged in this Rosen-Roback approach is one defined by location, rather than by the specific skills or sectors of workers: we think about the supply and demand for all workers in Detroit rather than the supply and demand for nurses or auto mechanics.

Let us first examine the labor demand curve more closely. Wherever people choose to live, they earn a wage from producing a good and then use that wage to buy goods and services. Let us assume that the wage they earn in any location i depends on two things: (1) the number of people living in that location; and (2) the productivities of that location. The result is a labor demand curve:

$$\ln w_i = \varepsilon^D \ln L_i + \ln C_i^D.$$

Figure 1
A Supply Shock in the Local Spatial Equilibrium



Source: Authors' creation.
Note: This figure illustrates the effect of an increase in the labor supply shifter on the equilibrium population and wages in a local spatial economy.

In this relationship, the terms for wages and quantity of labor are expressed in log terms, and so ε^D is the demand elasticity. C_i^D is the local productivity in region i that arises from its local geography. The local productivity may capture, for example, how productive the factors are in location i or the relative cost of capital in a location.

The elasticity of demand is typically assumed to be negative, such that the labor demand function in Figure 1 is downward sloping. The economic intuition behind this slope is often based on assuming decreasing returns to scale in production of the good or simply the presence of a fixed factor such as capital (for example, see Kline and Moretti 2014; Donaldson and Hornbeck 2016). In other words, there is diminishing marginal product for each additional unit of labor added in the location. Thus, as the population of a location increases, each additional worker is less and less productive, causing the wage to fall. But other scenarios are possible. For example, the presence of external economies also can affect the slope of the demand function. If more workers in a location result in everyone being more productive, the labor demand curve can become more elastic; if these external economies are sufficiently strong, the demand curve may even slope upwards. This situation can lead to outcomes like multiple equilibria or like "black hole" equilibria, where everyone lives in one location (for discussion, see Krugman 1991; Fujita, Krugman, and Venables 1999). While such scenarios have academic interest, in what follows we will stick with the more common (and, arguably, more empirically relevant) case of a downward-sloping demand curve.

If people each choose their place of residence to be as happy as possible, what makes people happy in this framework? Two things: higher consumption (so, all else equal, workers prefer higher real wages) and living somewhere nice (that is, a place with high amenities). In a model where everyone is identical, all inhabited locations must make people equally as happy. If prices are the same everywhere (so that the real wage is the nominal wage) and the amenity value of a location depends in part on how many other people live there, then workers' indifference across all inhabited locations generates this labor supply curve:

$$\ln w_i = \varepsilon^S \ln L_i - \ln C_i^S.$$

Again, the left-hand side of the equation is the wage for each worker in the region i, and L_i is the number of workers in the region. Because wages and the quantity of workers are expressed in logs, ε^S is the elasticity of labor supply. C_i^S is the local amenity in region i—for example, better parks or planetariums.[1]

Economists usually think of a supply curve as sloping upward, as the labor supply curve is shown in Figure 1. A common underlying assumption in this setting is that the supply curve will slope up as long as more people in a location make each individual less happy; for example, the presence of a housing market where a higher population drives up housing prices and rents or the existence of idiosyncratic preferences where a higher population means the marginal resident's match quality is worse can also lead to upward sloping labor supply curves.[2]

It is theoretically possible for the labor supply curve to slope downward (and issues of multiplicity and black holes to arise) if the amenity value of a location is increasing in its population, perhaps because of greater investments in public goods or greater variety in consumables in that specific location, but again, we will set that possibility aside here.

In this model, the equilibrium of economic activity—that is, the population and wage in a specific location—arises from combining the labor demand and labor supply curves. The spatial equilibrium is highlighted at point A in Figure 1.

[1] See the online Appendix A.1 for a particular microfoundation that delivers the specific labor demand and labor supply functions shown here.

[2] For a discussion of heterogeneous preferences and housing market, see Helpman (1998), Allen and Arkolakis (2014), Redding (2016), and Ahlfeldt et al. (2015).

To see how the "local" geography shapes the spatial equilibrium, consider a simple counterfactual scenario where the amenity value of residing in a location improved. For example, suppose the advent of air conditioning technology made the hot climate of the US Southwest less oppressive. An improvement in amenities shifts outward the labor supply curve, moving the equilibrium from point A to point B in Figure 1. The population in the location increases, but its wage declines. The US Southwest is now a better place to live, but the influx of workers depresses the wages.

The fact that we can analyze each location separately, depending on the amenity shock it receives, illustrates the somewhat paradoxical nature of the Rosen-Roback framework. It is a spatial model, but the distribution of economic activity depends only on local geography, not on what happens to other regions. Intuitive spatial features like where a location is located on a map and who its neighbors are entirely absent: it is a spatial model where space does not matter. By looking at one location at a time, it does not consider economic linkages between those locations.[3] Taking such linkages into account will create the concept of "global" geography which we introduce and analyze next.

The Role of Global Geography in the Spatial Distribution of Economic Activity

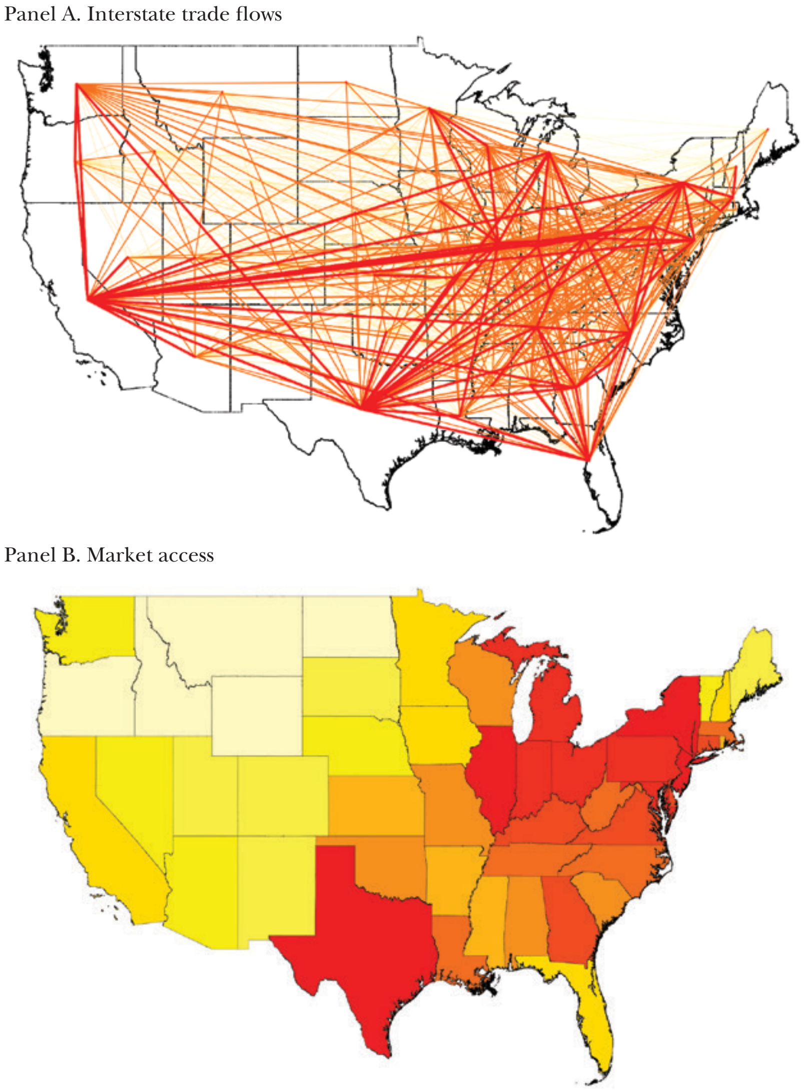
Different locations can be linked with each other in many ways: people may live in one location and work in another; people may migrate from one location; people may talk with each other, leading to the spatial diffusion of ideas; and so on. But perhaps the most obvious spatial linkage occurs through the flow of goods. Much of what an individual consumes is produced in another location: according to the 2017 United States Commodity Flow Survey (CFS 2017),[4] most freight shipments crossed state boundaries, with only 22 percent of the value of freight destined for a state also originating in the same state. Moreover, the pattern of trade flows are far from uniform. As panel A of Figure 2 highlights using the same data, nearby states trade more with each other while the total volume of trade increases with the size of the trading partners, a phenomenon originally observed in international trade flows and oftentimes referred to as "gravity" (Anderson 2010; Head and Mayer 2013).

How does incorporating such spatial linkages affect the spatial equilibrium? It turns out that much of the basic intuition above remains; in particular, we can still analyze the spatial equilibrium using the familiar techniques of supply and demand, albeit now augmented with a concept of both "local" and "global" geographies.

[3] In the Rosen-Roback framework, a change in the local geography in one location can have aggregate general equilibrium effects on, say, the price of capital. But such general equilibrium effects affect all locations equally and hence do not affect the spatial distribution of economic activity.

[4] The Commodity Flow Survey is conducted by the US government and is the primary source of data on within-US trade flows. In general, it is difficult to measure intracountry trade flows, making analysis of within-country trade difficult, although notable exceptions include work in Canada (Anderson and Van Wincoop 2003; McCallum 1995), India (Donaldson 2018), and the Philippines (Allen 2014).

Figure 2
Spatial Linkages and Market Access



Panel A. Interstate trade flows

Panel B. Market access

Source: Authors calculations based on data from CFS (2017).
Notes: This figure illustrates the spatial linkages across US states arising from trade flows. Panel A depicts the relative size of state-to-state bilateral trade flows, with thicker red lines indicating larger values and thinner yellow lines indicating smaller values. Panel B indicates the resulting (outward) market access of each state assuming trade costs T_{ij} are inversely proportional to distance, with the darker red states indicating greater outward market access and the lighter yellow states having lesser outward market access.

The Global Geography

The model discussed below is based on prior work (Allen and Arkolakis 2014), but variations of this spatial framework with equivalent or similar mathematical formulations have recently been used in a variety of frameworks.[5] The setup retains the same features as above, but now we introduce a key distinction: goods are no longer costlessly traded. There are trade relationships between different locations, governed by the presence of spatial frictions.

These spatial frictions can be described as the economic distance between regions i and j. Conceptually, economic distance is proportional to the value of trade flows between two locations (conditional on origin and destination fixed effects). There are many possible factors that influence the economic distance between locations—whether they speak the same language, share the same legal systems, share similar cultural heritages, and so on. But one of the most important contributors to economic distance is simply the geographic distance between any two locations. Indeed, one of the most robust empirical relationships in all of economics is that trade flows between locations are roughly inversely proportional to the geographic distance between them (for discussion, see Disdier and Head 2008; Chaney 2018). Put another way, a very good start to measuring "economic distance" is simply with geographic distance.

When spatial frictions exist and goods are no longer costlessly traded, two things change. First, the price of goods produced by workers in a location depends in part on how nearby the consumers of those products are. The closer the consumers are, the more demand for their products and the higher the price (and hence the higher the wage) that the workers can obtain. This outward market access affects the labor demand curve of a location. Second, the price of goods purchased by consumers in a location depends in part on how nearby the producers of those products are. The closer the producers, the lower the price for those products and the higher the real wage of the consumers. This inward market access acts as a shifter to the labor supply curve of a location.

Together, the outward and inward market accesses comprise the global geography of a location. Following Anderson and Van Wincoop (2003) and Redding and Venables (2004), the outward market access (MA_i^{out}) can be expressed algebraically as:

$$MA_i^{out} = \sum_j T_{ij} \times \frac{Y_j}{MA_j^{in}},$$

where T_{ij} is the inverse of economic distance between two locations and $Y_j = w_j L_j$ is the total income of location j. Intuitively, outward market access summarizes the selling potential of a market, indicating how well a region is connected to other locations. For example, New Jersey has a high outward market access because there are lots of potential consumers of its products in its neighboring states of New York and Pennsylvania. Outward market access is greater when neighboring locations are closer (that is, when the inverse economic distance T_{ij} is greater), which is especially

[5] See for example Redding (2016), Donaldson and Hornbeck (2016), Allen, Arkolakis, and Takahashi (2020), Faber and Gaubert (2019), and Eckert and Peters (2022). Redding and Rossi-Hansberg (2017) offer a comprehensive review of the quantitative spatial framework.

beneficial when those neighboring locations are richer (have higher Y_j) or have worse alternatives for buying their own goods (MA_j^{in}).

"Inward market access" is similarly defined as the capacity of locations to buy from other locations:

$$MA_j^{in} = \sum_i T_{ij} \times \frac{Y_i}{MA_i^{out}}.$$

For example, New Jersey also has high inward market access because it is able to purchase its goods from nearby large producers. Like outward market access, inward market access is greater the smaller the economic distance to other locations, and again this matters more when nearby locations either produce a lot higher (Y_i) or have poor alternatives for selling their goods (that is, have a lower MA_i^{out}).

Outward and inward market accesses are obviously quite closely related and, indeed, will be proportional to each other in the special case when economic distances are the same in both directions. Note, however, that the economic distance that matters for inward market access is the one in which a location is the destination, whereas for outward market access, the economic distance that matters is the one in which the location is the origin. As a result, when economic distances are not the same in both directions, the inward and outward market accesses will generally be different.

The global geography summarizes how each location depends on economic activity in all other locations, where closer locations are given greater weights. These algebraic formulations highlight that inward and outward market accesses are intertwined, with each dependent in part on the other. Despite this interdependence, it is straightforward to solve for both the market access measures as long one observes the income in each location and the economic distances between locations. The companion Matlab code available as an appendix to this paper provides a convenient algorithm for doing so.

Panel B of Figure 2 depicts the (outward) market access for each US state, where we proxy the inverse economic distance T_{ij} with the inverse of geographic distance, measured as the distance (as the crow flies) between the geographic center of each state. States with high economic output that are close to other states with high output such as those in the Northeast have good market access; states with less economic output that are far away from states with higher economic output such as Montana have poor market access. As we will discuss in the next main section, an appealing feature of this framework is that the inverse economic distance can also be measured more explicitly with a combination of observed bilateral trade flows and observed bilateral geographic characteristics such as distance or time of travel.

The Global Spatial Equilibrium

It turns out the global spatial equilibrium with spatial linkages can be analyzed using labor supply and demand curves, just as in the local spatial equilibrium above. Now, however, supply and demand will not only depend on local geography, but also on global geography. In particular, the labor demand now also depends on outward market access MA_i^{out}, becoming:

$$\ln w_i = \varepsilon_{local}^{D} \ln L_i + \varepsilon_{global}^{D} \ln MA_i^{out} + \ln C_i^D.$$

Better outward market access acts analogously to better local productivities, C_i^D, shifting the demand curve for local labor outwards with an elasticity ε_{global}^{D}. That elasticity is greater the less substitutable the goods produced in i are with goods produced elsewhere in the world.

Similarly, labor supply now depends on inward market access MA_i^{in}, becoming:

$$\ln w_i = \varepsilon_{local}^{S} \ln L_i + \varepsilon_{global}^{S} \ln MA_i^{in} + \ln C_i^S.$$

Better inward market access acts analogously to better local amenities C_i^S, shifting the supply curve for labor outwards with an elasticity ε_{global}^{S}, which again is larger the less substitutable goods produced in different locations are with each other.

The two limiting cases deserve special mention. When ε_{local}^{S} grows very large and approaches infinity, the local population becomes invariant to changes in economic conditions, whereas when ε_{local}^{S} becomes very small and approaches zero, labor supply is infinitely elastic to local economic conditions. These special cases correspond to important cases in the literature, as we will discuss below.

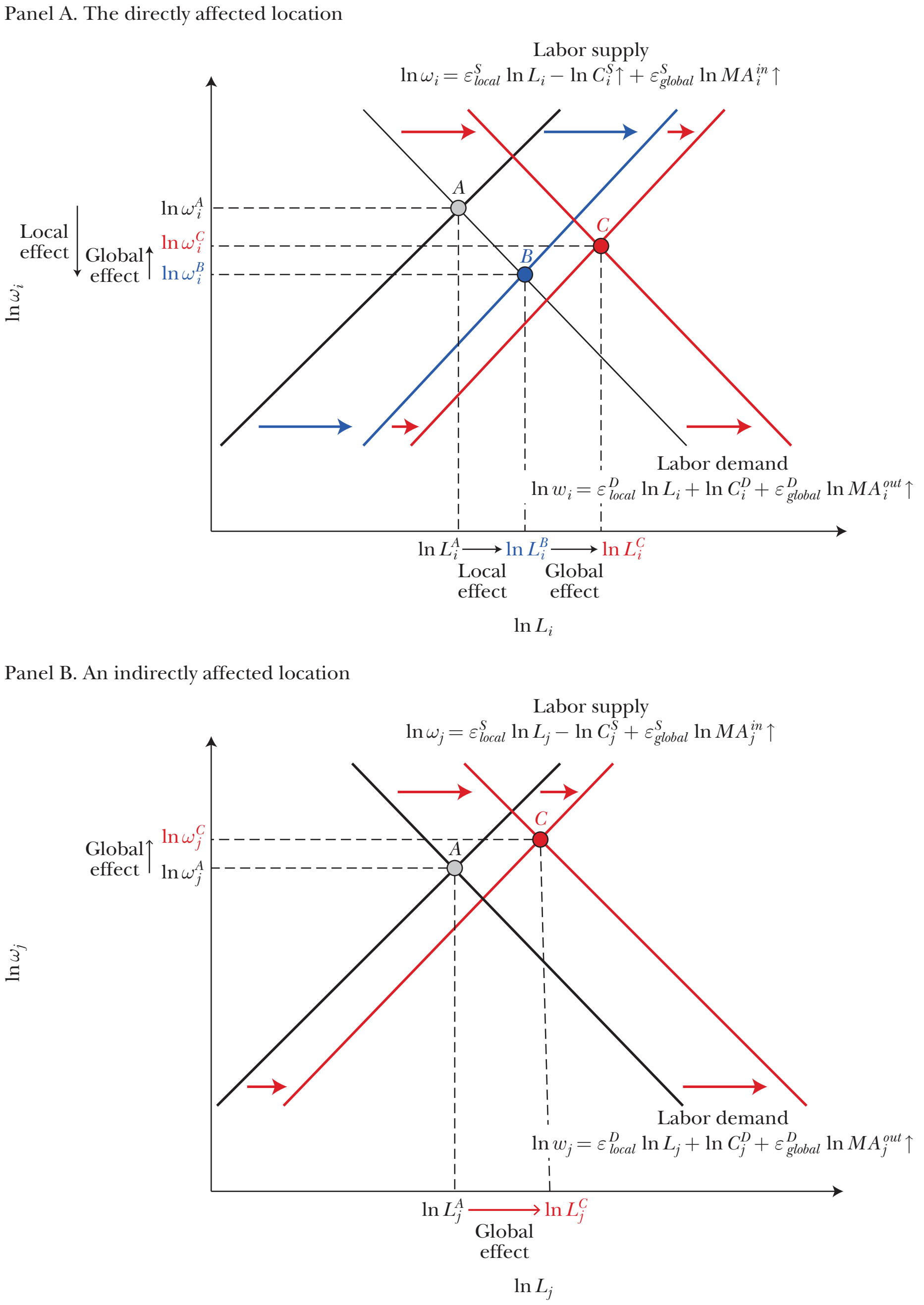
Given the global geography, the global spatial equilibrium is determined just as in the local spatial equilibrium above: find the wage and population in each location that equates supply with demand; point A on panel A of Figure 3 depicts such an equilibrium.

So what has changed in the global spatial equilibrium? The crucial insight is that the global geography in one location depends on the spatial equilibria in all other locations. If something changes about the local geography anywhere in the world, it will affect the global geography everywhere in the world, although it will affect nearby locations more than locations far away. Hence, the global geography puts space back into the spatial economy.

To illustrate this global spatial equilibrium, let us return to the example above. Suppose that air conditioning is invented, which makes some hot and previously inhospitable location i much more hospitable, raising the amenity of living there. Again, this innovation will shift outward labor supply curve in location i to point B in panel A of Figure 3, increasing the population in location i and reducing the wages. But the story does not end here, as this change in population and wages will affect the global geography. As long as the elasticity of local demand is greater than –1, the income Y_i of location i will increase, raising both the inward and outward market access and resulting in an additional shift outward to both the labor demand and labor supply curves. This additional global effect further increases the population in location i and mitigates the downward fall in wages, as illustrated in point C in panel A of Figure 3.

Figure 3

A Supply Shock in the Global Spatial Equilibrium



Panel A. The directly affected location

Labor supply
$\ln\omega_i = \varepsilon^S_{local}\ln L_i - \ln C^S_i\uparrow + \varepsilon^S_{global}\ln MA^{in}_i\uparrow$
A
$\ln\omega^A_i$
Local effect
Global effect
$\ln\omega^C_i$
C
B
$\ln\omega^B_i$
$\ln\omega_i$
Labor demand
$\ln w_i = \varepsilon^D_{local}\ln L_i + \ln C^D_i + \varepsilon^D_{global}\ln MA^{out}_i\uparrow$
$\ln L^A_i \longrightarrow \ln L^B_i \longrightarrow \ln L^C_i$
Local effect
Global effect
$\ln L_i$

Panel B. An indirectly affected location

Labor supply
$\ln\omega_j = \varepsilon^S_{local}\ln L_j - \ln C^S_j + \varepsilon^S_{global}\ln MA^{in}_j\uparrow$
C
Global effect
$\ln\omega^C_j$
A
$\ln\omega^A_j$
$\ln\omega_j$
Labor demand
$\ln w_j = \varepsilon^D_{local}\ln L_j + \ln C^D_j + \varepsilon^D_{global}\ln MA^{out}_j\uparrow$
$\ln L^A_j \longrightarrow \ln L^C_j$
Global effect
$\ln L_j$

Source: Authors' creation.
Note: This figure illustrates the effect of an increase in the labor supply shifter in one location its own equilibrium population and wages (panel A) and another neighboring location (panel B).

At the same time, changes in the economic activity in location *i* affect the global geography of other locations. Consider a neighboring location *j* initially in equilibrium, as illustrated by point A in panel B of Figure 3. Because the income of location its neighbor *i* has improved, both its supply and demand curves will shift outwards as well. Intuitively, the greater nearby economic activity both increases the demand for the goods produced in *j* and increases the supply of goods consumed in *j*. As a result, the population in *j* increases too (and its wages rise), changing its equilibrium to point C in panel B of Figure 3, despite there being no change in its own local geography.[6]

But will changes in the economic activity in location *j* not have subsequent impacts on the global geography in all other locations? And will those changes not have even further impacts on the global geography, ad infinitum? Yes and yes: indeed, this infinite feedback loop between the global geography in every location is part of what makes the global spatial equilibrium so interesting to study. In reality, point C in panels A and B of Figure 3 represents the limit of the infinite sequence of these adjustments of each location's global geography to adjustments made in the global geography everywhere else. Indeed, this iterative process is what both the algorithm for calculating the equilibrium change in market accesses in the companion Matlab code and many tools for studying the mathematical properties of the equilibrium system are based upon.[7]

Having shown how one can determine the global spatial equilibrium through the use of supply and demand curves, we now turn to describing the process through which this framework can be combined with spatial data to assess the impact of changes in geography on the real world spatial distribution of economic activity.

Estimating Labor Supply and Demand

In the previous section, we saw how a supply and demand framework can be used to understand how changes in the geography affect the distribution of economic activity across spatially connected locations. One of the most attractive

[6] Whether nominal wages rise or fall—that is, whether outward or inward market access increases more—depends on the choice of the numeraire. Here we set mean wages equal to one as the numeraire, so falling wages in location *i* must be offset by rising wages elsewhere.

[7] In the special case where the augmented labor supply curve is infinitely elastic, the local and global demand elasticities are equal in magnitude, and the inverse economic distances are symmetric, the equilibrium global economy is one in which the wages and populations of each location are (log) proportional to the eigenvector centrality of a location in the network defined by the world geography (that is, by the combination of the economic distances, productivities, and amenities). Higher eigenvector centrality means that a node in a network is nearby to other nodes with high eigenvector centralities. Here, it means that locations are more populated (and wealthier) the closer they are to other more populated (and wealthy) locations. Moreover, the eigenvalue of the system corresponding to this eigenvector turns out to be the welfare of the global economy (which is characterized by a single scalar because the infinitely elastic labor supply ensures welfare is equalized across all locations). In the more general case, the equilibrium of the spatial economy constitutes a network system of nonlinear equations. The properties of such systems remains an active field of research: Allen, Arkolakis, and Li (2020) offer a starting point.

aspects of the global spatial frame work described above is its ability to integrate seamlessly with readily available spatial data. In this section, we describe this interplay between theory and data.

Spatial Economic Data: Local and Linkages

We focus here on two types of spatial data: data on the local economic activity of a location and data on the strength of economics linkages between locations across space.

Suppose that a researcher can observe in the data how many people reside in a certain location L_i and the total income of a location Y_i. Indeed, such data are readily available; for example, in the United States, population data and income data at the county level can be constructed from the decennial Census going back to the year 1840. The IPUMS (Integrated Public Use Microdata Series) National Historical Geographic Information Systems (Manson 2020) has provided an enormous public good in assembling these data and making them publicly available. Even in parts of the globe where spatially disaggregated income data are not readily available, one can proxy for economic activity using satellite data on the intensity of lights at nighttime, a practice pioneered by Henderson, Storeygard, and Weil (2012) and summarized in this journal in Donaldson and Storeygard (2016). Furthermore, databases that assemble information from various sources provide disaggregated information on economic activity at a granular geographic level, such as the G-econ database (Nordhaus and Chen 2006) that provides proxies of global income and population at the one-arc degree.

We furthermore assume that all income accrues to labor, which allows us to recover average wages for a location given knowledge of income and population. This strong assumption clearly abstracts from sources of income like capital, landholdings, firm profits, and others. One could argue that all these sources of income eventually accrue to individuals as well; indeed, as long as the income remains in a particular location, the predictions of the global spatial framework does not change by incorporating these other sources of income. (For example, as long as individuals in a location own their own homes, a model where individuals spend money on housing is no different—we say it is "isomorphic"—to the framework described above.) But in reality, not all income earned in a location accrues to the labor in that location, and such spatial flows of income would present another linkage between locations that we abstract from here.

Next consider data on economic linkages across space. As noted earlier, geographic distance is offers a convenient proxy for economic distance. But recently, researchers have begun to improve upon the distance proxy with measures of actual travel costs between locations. For example, Donaldson (2018) estimates the relative cost of traveling between locations by means of road, rail, and waterways by calculating the lowest cost route using Dijkstra's (1959) algorithm—the same algorithm used by, for example, Google Maps. Allen and Arkolakis (2014) use a continuous space extension of the Dijkstra algorithm known as the Fast Marching Method (Tsitsiklis 1995; Sethian 1999) to calculate travel times along the optimal route between locations. Allen and Arkolakis (2022) offer an analytical solution for the inverse economic distance as a function of the underlying transportation network.

Intuitively, these related approaches all share two advantages. First, they provide more precise estimates of the economic distance between two locations than distance alone would provide. (For example, Milwaukee, Wisconsin, and Grand Rapids, Michigan, are about 115 miles apart as the crow flies, but travel between the two around Lake Michigan more than doubles the distance). Second, accounting for the underlying transportation network allows researchers to assess how changes in transportation infrastructure (for example, improving the interstates I-90 and I-94 that connect Milwaukee and Grand Rapids) affect the spatial distribution of economic activity.

For any observed measure(s) of the economic linkages, the inverse economic distance (T_{ij}) can then be constructed by regressing the observed (log) value of trade flows on those measures, conditioning on the origin and destination fixed effects. The predicted values of this gravity-model regression (excluding the estimated fixed effects) are the implied inverse economic distance.[8] For example, if one uses travel times as a measure of economic linkages, the inverse economic distance would be the product of travel time and its estimated coefficient from such a regression.

Estimating Supply and Demand

Given measures of income Y_i in each location and a measure of the strength of the linkages T_{ij} between locations, we can calculate the global geography of every location—that is, the inward and outward market accesses MA_j^{in} and MA_i^{out}.[9] We provide an iterative algorithm for solving that nonlinear system of equations in the companion Matlab code.

Now let us return to our augmented supply and demand equations for the global case. We observe the left-hand-side price variable, the wage for each location w_i, and the right-hand-side quantity variable, the population L_i We also observe the data needed to calculate the market access variables (MA_i^{in} and MA_j^{out}).

We would like to estimate the coefficients on the right hand side variables, which represent the local and global elasticities of supply and demand. In doing so, the residual terms will be equal to measures of local productivity and local amenities

[8] An alternative procedure would be to calibrate the inverse economic distance to exactly match the observed bilateral trade flows by including the regression residual in its construction. Such a procedure—which is closely related to the "exact hat algebra" pioneered by Dekle, Eaton, and Kortum (2008) and discussed in Costinot and Rodríguez-Clare (2014)—can result in an over-fitting problem when conducting counterfactuals (Dingel and Tintelnot 2020).

[9] Recovering the global geography from the observed income and economic distances is a well-behaved problem. One can show using tools from Allen, Arkolakis, and Li (2020) that there exists unique (to-scale) inward and outward market accesses MA_j^{in} and MA_i^{out} that solve the equations for any set of incomes Y_i and inverse economic distances T_{ij}.

(that is, ln C_i^D and ln C_i^S).[10] Or put another way, we would like to estimate a system of supply and demand curves where we observe data on equilibrium outcomes of price and quantity at different times, which poses problems that are all-too-well understood!

How do we go about estimating our supply and demand curves? It might perhaps be more informative to start with what not to do. Following in the footsteps of Baldwin and Taglioni (2006), let us award medals for different types of errors that can arise, ranking them from most to least obvious.

The Bronze Medal Error

One glaring mistake in estimating supply and demand equations and—our "bronze medal" error—would be to use ordinary least squares regression. This approach is clearly not appropriate due to familiar simultaneity issues: what appears in data on wages and workers are the intersections of supply and demand curves, which do not trace out the shape of either a supply or a demand curve, but rather a series of movements in both of them. (To put it another way, because the right-hand-side population variable is determined in equilibrium from equating supply and demand, it will be correlated with both the productivity and amenity shifters.) As a result, the coefficient from such an ordinary least squares regression will not recover either the supply or demand elasticity.

One strategy for overcoming this bronze medal error would be to employ instrumental variables; for example, using variation in the amenity ln C_i^S as an instrument for the equilibrium population to estimate the labor demand elasticity and using variation in the productivity as an instrument for the equilibrium population to estimate the supply elasticity. Conceptually, this involves looking at a source of shifts in labor supply (in this case, local amenities) to trace out a labor demand curve, and a source of shifts in labor demand (in this case, changes in local productivity) to trace out a labor supply curve. As long as the chosen instrumental variation in the amenities and productivities are uncorrelated, this will yield consistent estimates of the demand and supply elasticities.

What are examples of such instruments? One example comes from Glaeser and Gottlieb (2009), who argue that the advent of air conditioning improved the amenity of locations with warm climates. Under the assumption that the climate of a location is not also correlated with the change in the productivity of a location, the climate of a location can be used as an instrument for change in population to identify the demand elasticity (for example, Allen and Donaldson 2020).

Conversely, Allen and Donaldson (2020), following Bustos, Caprettini, and Ponticelli (2016), argue that increased global demand for soy improved the productivity of locations particularly well-suited for the production of soy. Under the assumption that the potential yield of soy in a location (say, relative to its potential

[10] This approach of recovering the underlying geography based on the supply and demand residuals is equivalent (but perhaps easier to digest) to an approach that directly inverts the equilibrium market clearing conditions, as in Allen and Arkolakis (2014) and Redding (2016).

yield for corn) did not also change the amenity of a location, the potential relative yield of soy to corn can be used as an instrument to identify the supply elasticity. Of course, the climate or agroclimatic properties are likely correlated with myriad characteristics of a location, making it unlikely these assumptions hold when comparing wages and populations across locations in cross section at a point in time. As such, it is preferable to rely on panel variation, looking at changes in wages and populations across locations over time (or, equivalently, including location fixed effects in the estimation of the supply and demand equations).

The Silver Medal Error

Somewhat less obviously, our "silver medal" error would be to ignore the spatial linkages between locations and simply estimate supply and demand using the local supply and demand equations based on the Rosen-Roback model. However, doing so ignores the variation in inward and outward market access across locations, relegating that variation to the residual term.

The instrumental variable strategy just described to address simultaneity bias is insufficient to address this bias. To see this, suppose you are estimating the labor demand equation, while using an amenity shifter like the arrival of air conditioning as an instrumental variable for population. Even if that amenity shifter is uncorrelated with productivities, it will be correlated with outward market access, biasing the estimate of the demand elasticity. Indeed, the only situation where this bias does not arise is in the special case when all locations share the same market access (as in the local spatial equilibrium).[11]

Fortunately, avoiding this mistake is straightforward: from the discussion above, one can construct measures of inward and outward market access from readily available spatial economic data. Including these market access measures in the supply and demand equations is a simple remedy to avoid the silver medal error.

The Gold Medal Error

An even more subtle concern is that outward and inward market access measures are themselves almost surely correlated with the productivity and amenity of a location. After all, the market access of a location depends in part on its own economic activity, which of course depends in equilibrium on its productivity and amenity. As a result, just including the market access measures in the supply and demand equations as controls will result not only in biased estimates of both the local and the global elasticities of supply and demand.

To address this concern, one can again use an instrumental variables strategy, instrumenting for both the population in a location and for the market access of

[11] Our "silver medal" error is similar in spirit to Baldwin and Taglioni's (2006) "gold medal" error of failing to control for variation in market access in gravity equations. The two errors are distinct because unlike a gravity regression, the supply and demand regressions are not estimated using bilateral flows. As a result, their proposed solution of controlling for market access with origin and destination fixed effects does not apply here.

that location. We discussed above possible instruments for the population; what about for market access? An appropriate instrument would be correlated with market access, but uncorrelated with local productivities or amenities.

In conceptual terms, think of market access as a type of inverse economic distance-weighted average of economic activity near a location. For an appropriate instrumental variable, suppose you use the measures of local productivities and amenities along with plausible values of the model elasticities to calculate the local equilibrium of a hypothetical economy using the basic local-area Rosen and Roback supply and demand equations. In this hypothetical economy, spatial linkages do not matter and the only heterogeneity in productivities and amenities across locations arise from observables. Next, combine the implied equilibrium income in each location from this hypothetical economy with the observed economic distance and use the market access expressions above to calculate what the market access would be in such a hypothetical economy. This hypothetical market access measures how well connected each location is to the rest of the world, if the income in each location depended only on its observed productivities and amenities.

The hypothetical market access is a valid instrument for the actual market access under the assumption that observed productivities and amenities elsewhere in the world are uncorrelated with a location's own unobserved productivities and amenities. Using the hypothetical market access as an instrument then isolates the impact of market access on the supply and demand curves using this variation in productivities and amenities elsewhere through the spatial structure of the model.[12] Examples of such "model implied" instruments can be found in Monte, Redding, and Rossi-Hansberg (2018), Allen, Arkolakis, and Takahashi (2020), and Adão, Arkolakis, and Esposito (2019).

Taking Stock

Suppose you have successfully avoided the bronze, silver, and gold medal errors by estimating the labor supply and demand curves while appropriately using instrumental variables for the observed population and the market access terms. Now what?

You are now armed with estimates of the model elasticities, data on wages, populations, and market access terms, and with residuals terms from the supply and demand equations that correspond to the productivities and amenities in each location. Put another way, if you know the supply and demand elasticities, you can always find the local geography such that the observed distribution of economic activity—combined with the inverse economic distances you have constructed—is the global spatial equilibrium of the model.

[12] Another possibility would be to construct an instrument based on the augmented global supply and demand equations but excluding the own location (and perhaps also nearby locations) from the sum. Even if there is no spatial correlation in the productivity and amenity of locations, however, the equilibrium economic activity elsewhere depends in part on the economic activity of the own location (and hence the own productivity and amenity shifters), so such an instrument is unlikely to satisfy the exclusion restrictions.

Because you have recovered the geography that is consistent with the observed economic activity and you know the model elasticities, you are now able to assess how changes to the geography will affect the global spatial equilibrium. In the next section, we will discuss ways in which this approach can inform understanding concerning the effects of various events and policy decisions.

Understanding the Spatial Impact of Economic Policies

We have seen how the global and local geographies interact through supply and demand to shape the spatial equilibrium and how those supply and demand curves can be combined with spatial data to apply the framework to the real world. Now we are equipped to describe the many types of questions that can be addressed with such a framework. We classify these questions into three types: those examining the impact of changes to the local geography, those examining the impact of changes to the global geography, and those which extend the framework above to incorporate additional spatial linkages beyond the flow of goods. We make no pretense here of offering a full survey of the literature; instead, our goal is to illustrate the extraordinary range of this work across events, policies, places, and times.

Local Geography Shocks

Consider first the question of how changes to local geography—changes to amenities which shift the supply curve or changes to productivities which shift the demand curve—affect the spatial distribution of economic activity.

Changes in the natural environment due to climate change offer many such examples. Rising sea levels and the resultant flooding both reduce the amount of land available for production and reduce the attractiveness of living in a coastal location, shifting both supply and demand curves in such locations inward, inducing populations to migrate elsewhere. Desmet et al. (2018) study the long-run impact of coastal flooding using a dynamic variation of the framework described here, finding that approximately 1.5 percent of the world population will be displaced by the year 2200 under current projections of the extent of flooding. Changing temperatures and patterns of precipitation also affect the suitability of different locations for producing different types of crops, affecting the productivity of different locations. Costinot, Donaldson, and Smith (2016) examine the long-run impact of estimated future changes in agricultural productivity across the globe to assess its impact on the spatial distribution of economic activity, estimating that climate change will reduce the global value of agricultural output by approximately one-sixth.

Conflict and war can also reduce local productivities and amenities, although it remains an outstanding question for how long after the conflict these effects persist. For example, Davis and Weinstein (2002) examine the rebuilding of Japan after World War II, finding that the postwar distribution of economic activity closely mirrored the pre-war distribution, suggesting that wartime destruction was not enough to overcome fundamental characteristics of different locations. In contrast, Chiovelli, Michalopoulos, and Papaioannou (2018) find the removal of landmines in the period after Mozambique's civil war had substantial impacts on the spatial distribution of economic activity, especially after accounting for the impacts of the de-mining on market access—that is, on the global geography.

Technological innovations may also increase the productivities in certain locations, shifting the labor demand curve outward. For example, Bustos, Caprettini, and Ponticelli (2016) present evidence that the introduction of genetically modified soybeans in Brazil had heterogeneous effects across areas with different soil and weather characteristics, and also was a labor-saving technology that ended up boosting industry. Caliendo et al. (2018) extend the framework above to incorporate intersectoral linkages along with spatial linkages to examine, for example, how local productivity improvements resulting from California's computer industry boom and the introduction of shale oil production in North Dakota affected the spatial distribution of economic activity. Some interesting topics for future research along these lines include the spatial effects of automation (as in Acemoglu and Restrepo 2020) or new technologies that allow for remote work (as in Dingel and Neiman 2020; Althoff et al. 2022).

Place-based policies enacted by the government can also be viewed as shifts to the local demand or supply curves (depending on the particular nature of the policy). For example, Diamond and McQuade (2019) show that tax credits for low-income housing projects across 129 counties nationwide raised housing prices and reduced crime rates in low-income neighborhoods, but reduced housing prices in high-income neighborhoods. Some recent work seeks to characterize the trade-offs of such policies; for example, how policies that attract high-skill workers to low-wage cities can have broader social benefits and the equity-efficiency trade-offs of focusing place-based policies on locations with a dense concentration of low-income households (for discussion, see Fajgelbaum and Gaubert 2018; Gaubert, Kline, and Yagan 2021).

Global Geography Shocks

Now let us turn our attention to how changes to global geography—changes in the economic distances and the resulting changes in the market access—affect the spatial distribution of economic activity.

Investment in transportation infrastructure which reduces the economic distance between locations is a natural application for evaluating changes to global geography. For example, the US interstate highway system increased US welfare by 1.0 to 1.4 percent of GDP, more than its costs (Allen and Arkolakis 2014); the US railroad system constructed in the second half of the nineteenth century more than doubled the price of land in nearby agricultural counties (Donaldson and Hornbeck 2016); the Los Angeles Metro rail system increased commuting, but with little effect on productivity or amenities, and thus has considerably larger costs than benefits (Severen 2021); the Appalachian Development Highway System started in

1965 did benefit Appalachian counties, but most of the benefits accrued outside the region (Jaworski and Kitchens 2019); and the arrival of the steam railway in mid-nineteenth century London led to a doubling of population and land prices, as well as a geographical separation of workplaces and residences (Heblich, Redding, and Sturm 2020). Recent work has also examined the distributional implications of such infrastructure investments; for example, transportation infrastructure investments in New York City from 1870 to 1940 seem to have caused greater racial sorting and disparities (Lee 2022) and the recently constructed national highway system in China benefits the economy of larger regional cities at the expense of rural regions (Baum-Snow et al. 2020).

While the basic framework above abstracts from the possibility that the economic distances may depend in part on the amount of trade between two locations, Duranton and Turner (2011) demonstrate the empirical relevance of congestion by showing that neither additional roads nor mass transit seem to reduce congestion in US cities. Recent work has made substantial progress incorporating congestion into spatial frameworks like the one described above. For example, Fajgelbaum and Schaal (2020) study optimal transportation networks in the presence of traffic congestion. In applying their framework to European countries, they find that the desirable network depends on whether they focus on flows within countries or flows between countries. In a similar spirit, Allen and Arkolakis (2022) develop a spatial framework that includes congestion and apply it the US highway network and the Seattle road network. These types of frameworks could also be used to evaluate congestion imposing tolls in specific areas of the cities, such as the London or Singaporean traffic toll system or the congestion price system suggested for downtown Manhattan.

Other recent work has sought to consider congestion in the context of ports, sea routes, and supply chains. In particular, the Allen and Arkolakis (2022) spatial framework for transportation and congestion has been applied to study the effect of several recent events in global shipping on the distribution of economic activity. For example, the 2016 expansion of the Panama Canal expanded trade between pairs of countries using the canal by 9 to 10 percent, although the costs of the expansion were borne by Panama (Heiland et al. 2019); the expansion of container shipping and Chinese-financed development of seaports across Africa and Asia is leading to reallocations away from more expensive ports like Singapore (Ducruet et al. 2020); and entrepots, defined as shipping hubs that serve an intermediate role between place of origin and destination, play a key role in holding down global shipping costs (Ganapati, Wong, and Ziv 2020).

Another branch of this work looks at intermodal shipping: for example, how the construction of expressways in China early in the twenty-first century boosted exports (Fan, Lu, and Luo 2021) and how to identify the nodes between road, rail, and ports in the US economy that would provide the greatest gains from additional investment (Fuchs and Wong 2022). An exciting new area of work builds on the approach of Brancaccio, Kalouptsidi, and Papageorgiou (2020), who develop a model of endogenous route choices of exporters and endogenous transportation costs to study the global bulk shipping that constitutes 80 percent of world trade and evaluate the effect of large infrastructure projects such as the expansion of the Panama canal. Conwell (2022) combines endogenous route choices and traffic to find that an optimal subsidy on minibus entry in Cape Town, South Africa, may particularly benefit low-skill workers on long routes.

A classic example of changes in global geography arises from changes in international trade policy, like changes in tariffs. For example, Topalova (2010) examines the impact of the 1991 Indian tariff reduction to measure the impact of trade liberalization on poverty and rural districts, in which production sectors more exposed to tariff declines experienced slower decline in poverty and lower consumption growth. The recent escalation of tariff measures by large economies such as the United States and China has generated a renewed interest on the impact of tariff increases on the spatial distribution of economic activity, following the influential work of Fajgelbaum et al. (2020), who find that the recent US-China trade war reduced US real income by $7.2 billion, with the benefits of tariffs concentrated in politically competitive counties.

A final set of questions can be thought of as how changes to the local geography in some locations affect the economy elsewhere through the global geography. For example, beginning with the influential work of Autor, Dorn, and Hanson (2013), there has been much work on how productivity increases in China have affected workers in the United States and elsewhere through spatial linkages. Autor, Dorn, and Hanson (2013) found that US labor markets that previously included import-competing manufacturing industries experienced job and economic losses from the "rise of China." Caliendo, Dvorkin, and Parro (2019) use a spatial framework like the one above (expanded to include multiple sectors) to conclude that while there was an overall loss of manufacturing jobs from the rise of China, the US economy as a whole benefited, albeit with considerable variation across sector-state labor markets. The increase in demand elsewhere for goods or services in a location provides another example: Faber and Gaubert (2019) show that increasing international demand for tourism in Mexico causes large and significant local economic gains, which are in part driven by positive spillovers on manufacturing. In contrast, Allen et al. (2021) find that increasing international demand for tourism in Barcelona reduces the welfare of many local residents by increasing prices and crowding out local consumption.

Alternative Spatial Linkages

The framework developed above focuses on spatial linkages between locations that arise through the trade of goods. But of course people interact across space in many ways, including commuting, migration, or even social and business personal networks (for example, Christakis and Fowler 2009). Some recent advances have incorporated other types of interactions into spatial frameworks like the one developed here.

Following the seminal work of Ahlfeldt et al. (2015), which considered how the rise and fall of the Berlin Wall affected the spatial distribution of economic activity in that city, a number of papers have examined the impact of spatial interactions that arise through commuting flows. For example, Severen (2021), mentioned earlier, separates the commuting effect of the Los Angeles Metro from productivity or amenity effects, while Zárate (2022) find that extensions of subway lines in Mexico City lead to increased commuting and a shift from informal to formal jobs. Monte, Redding, and Rossi-Hansberg (2018) and Allen, Arkolakis, and Li (2015) combine commuting and spatial linkages in a single model: the first study finds that communities which win a competition for location of large plants have greater benefits if they have a more open commuting network; the second considers optimal zoning policy and finds Chicago would benefit from having more residences downtown and more business activity in outlying neighborhoods.

A related literature incorporates spatial linkages arising through altered migration patterns, extending the framework above to a dynamic setting. While the steady state (or balanced growth path) of these models resemble the static framework above, they are also able to yield predictions on the time it takes the economy to adjust to changes in geography. For example, in a global model with realistic geography, Desmet, Nagy, and Rossi-Hansberg (2018) examine different scenarios for migration and how eliminating migration restrictions could triple global welfare. Allen, de Castro Dobbin, and Morten (2018) show that walls built along the US-Mexico border altered migration patterns between Mexican municipalities and US counties. Tombe and Zhu (2019) argue that declining costs of internal migration in China can account for one-third of the aggregate growth in China's labor productivity from 2001 to 2005. Peters (2022) finds that the expulsion of ethnic Germans from eastern Europe after World War II, and their return to West Germany, increased aggregate income per capita by about 12 percent after 25 years. Finally, Kleinman, Liu, and Redding (2021) find that the interaction of migration and capital investment can help to explain why convergence of incomes between US states declined between 1965 to 2015.

Another spatial linkage garnering recent attention is the formation of production linkages across firms. For example, lower costs of searching for and creating linkage between heterogeneous buyers and sellers can drive down marginal costs, as Bernard, Moxnes, and Ulltveit-Moe (2018) and Bernard, Moxnes, and Saito (2019) find in applying their models to improved flow of people in Japan and to Norwegian customs data. Yet another spatial linkage can be measured by taking advantage of new data sets to assess the role of knowledge diffusion. Using nationally representative smartphone data, Couture et al. (2020) examine patterns of travel and communication. While using highly granular smartphone data, Atkin, Chen, and Popov (2022) find substantial returns to what are actually face-to-face interactions in Silicon Valley. Using Facebook data grouped by zip code (and thus anonymized), Chetty et al. (2022a, b) look at personal connections across socioeconomic groups and within cliques to study associations with economic mobility and determinants of connectedness.

Related studies look at the effects of new information technologies, documenting how the spatial spread of information can affect the distribution of economic activity. For example, Steinwender (2018) finds that the introduction of the trans-Atlantic telegraph in 1866 provided information that affected cotton prices and trade flows, with gains equivalent to 8 percent of export value. Allen (2014) shows that including information frictions can make sense of observed patterns of regional agricultural trade flows prices in the Philippines. Akerman, Leuven, and Mogstad (2022) find that this improved access to information in makes trade patterns more sensitive to distance and economic size using broadband expansion in Norway.

Recent research has incorporated even more types of spatial linkages including electricity transmission (Arkolakis and Walsh 2022), piped water (Coury et al. 2022), and natural gas pipelines (Bachmann et al. 2022). The possibilities of adding additional spatial linkages or combining multiple types (or multiple layers) of linkages seem limitless. Moreover, extending the framework to include such interactions brings more realism and helps to illuminate the many ways in which geography shapes the spatial economy.

Conclusion

This article has sought to serve three purposes. First, it was meant as an introduction to the reader about how geography shapes the spatial distribution of economic activity. In the classic Rosen-Roback framework, the answer depends solely on the "local" geography of each location and the equilibrium spatial distribution can be determined through familiar analysis of supply and demand curves. The major innovation of the new generation of economic geography models is to incorporate the spatial linkages between locations—putting space into the spatial model. The equilibrium can continue to be understood using the same supply and demand curves, but is appropriately augmented to incorporate the impacts of the "global" geography.

The second purpose was to guide the reader through the process of combining these spatial models with spatial data to understand how geography shapes the real world spatial economy. Detailed spatial data are now readily available and researchers can apply these data to the theory using the well-understood process of estimating supply and demand curves. With spatial linkages between locations arise potential pitfalls in estimation, but we offer strategies for traversing such issues. The end result is the ability to recover the underlying local and global geography such that the theory and data exactly correspond, allowing a researcher the ability to assess the impacts of any change in geography on the real world spatial distribution of economic activity.

Finally, we demonstrate the power of this close marriage between theory and data by highlighting the many types of questions that can be addressed. The types of questions and topics that can be examined using the framework here spans an incredibly wide range of topics, such as economic history, environmental, labor, public finance, urban, and international topics, to name a few. This is an exciting time to be working on spatial issues: we have a new set of tools applicable to many interesting questions, most of which have yet to be tackled.

■ *We thank our editors Erik Hurst, Nina Pavcnik, Timothy Taylor, and Heidi Williams as well as Steve Redding, Anson Soderbery, and Sayantan Roy for exceptionally helpful feedback and Saptarshi Majumdar for excellent research assistance. The authors acknowledge support by the National Science Foundation under grants SES-1658838 and SES-1658875.*

References

Acemoglu, Daron, and Pascual Restrepo. 2020. “Robots and Jobs: Evidence from US Labor Markets.” *Journal of Political Economy* 128 (6): 2188–2244.

Adão, Rodrigo, Costas Arkolakis, and Federico Esposito. 2019. “General Equilibrium Effects in Space: Theory and Measurement.” NBER Working Paper 25544.

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. “The Economics of Density: Evidence from the Berlin Wall.” *Econometrica* 83 (6): 2127–89.

Akerman, Anders, Edwin Leuven, and Maren Mogstad. 2022. “Information Frictions, Internet, and the Relationship between Distance and Trade.” American Economic Journal: Applied Economics 14 (1): 133–63.

Allen, Treb. 2014. “Information Frictions in Trade.” *Econometrica* 82 (6): 2041–83.

Allen, Treb, and Costas Arkolakis. 2014. “Trade and the Topography of the Spatial Economy.” Quarterly Journal of Economics 129 (3): 1085–1140.

Allen, Treb, and Costas Arkolakis. 2022. “The Welfare Effects of Transportation Infrastructure Improvements.” *Review of Economic Studies* 89 (6): 2911–57.

Allen, Treb, and Costas Arkolakis. 2023. “Replication data for: Economic Activity across Space: A Supply and Demand Approach.” American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E186022V1.

Allen, Treb, Costas Arkolakis, and Xiangliang Li. 2015. “Optimal City Structure.” Unpublished.

Allen, Treb, Costas Arkolakis, and Xiangliang Li. 2020. “On the Equilibrium Properties of Network Models with Heterogeneous Agents.” NBER Working Paper 27837.

Allen, Treb, Costas Arkolakis, and Yuta Takahashi. 2020. “Universal Gravity.” Journal of Political Economy 128 (2): 393–433.

Allen, Treb, Cauê de Castro Dobbin, and Melanie Morten. 2018. “Border Walls.” NBER Working Paper 25267.

Allen, Treb, and Dave Donaldson. 2020. “Persistence and Path Dependence in the Spatial Economy.” NBER Working Paper 28059.

Allen, Treb, Simon Fuchs, Sharat Ganapati, Alberto Graziano, Rocio Madera, and Judit Montoriol-Garriga. 2021. “Urban Welfare: Tourism in Barcelona.” Unpublished.

Althoff, Lukas, Fabian Eckert, Sharat Ganapati, and Conor Walsh. 2022. “The Geography of Remote Work.” *Regional Science and Urban Economics* 93: 103770.

Anderson, James E. 2010. “The Gravity Model.” *Annual Review of Economics* 3: 133–60.

Anderson, James E., and Eric van Wincoop. 2003. “Gravity with Gravitas: A Solution to the Border Puzzle.” *American Economic Review* 93 (1): 170–92.

Arkolakis, Costas, and Conor Walsh. 2022. “Clean Growth.” Unpublished.

Armington, Paul S. 1969. “A Theory of Demand for Products Distinguished by Place of Production.” IMF

Staff Papers 16 (1): 159–78.

Atkin, David, M. Keith Chen, and Anton Popov. 2022. "The Returns to Face-to-Face Interactions: Knowledge Spillovers in Silicon Valley." NBER Working Paper 30147.

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

Bachmann, Rüdiger, Daivd Baqaee, Christian Bayer, Moritz Kuhn, Andreas Löschel, Benjamin Moll, Andreas Peichl, Karen Pittel, and Moritz Schularick. 2022 "What If? The Economic Effects for Germany of a Stop of Energy Imports from Russia." ECONtribute Policy Brief 028.

Baldwin, Richard, and Daria Taglioni. 2006. "Gravity for Dummies and Dummies for Gravity Equations." NBER Working Paper 12516.

Baum-Snow, Nathaniel, J. Vernon Henderson, Matthew A. Turner, Qinghua Zhang, and Loren Brandt. 2020. "Does Investment in National Highways Help or Hurt Hinterland City Growth?" *Journal of Urban Economics* 115: 103124.

Bernard, Andrew B., Andreas Moxnes, and Yukiko U. Saito. 2019. "Production Networks, Geography, and Firm Performance." Journal of Political Economy 127 (2): 639–88.

Bernard, Andrew B., Andreas Moxnes, and Karen Helene Ulltveit-Moe. 2018. "Two-Sided Heterogeneity and Trade." Review of Economics and Statistics 100 (3): 424–39.

Brancaccio, Giulia, Myrto Kalouptsidi, and Theodore Papageorgiou. 2020. "Geography, Transportation, and Endogenous Trade Costs." *Econometrica* 88 (2): 657–91.

Bustos, Paula, Bruno Caprettini, and Jacopo Ponticelli. 2016. "Agricultural Productivity and Structural Transformation: Evidence from Brazil." *American Economic Review* 106 (6): 1320–65.

Caliendo, Lorenzo, Maximiliano Dvorkin, and Fernando Parro. 2019. "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock." *Econometrica* 87 (3): 741–835.

Caliendo, Lorenzo, Fernando Parro, Esteban Rossi-Hansberg, and Pierre-Daniel Sarte. 2018. "The Impact of Regional and Sectoral Productivity Changes on the U.S. Economy." *Review of Economic Studies* 85 (4): 2042–96.

Chaney, Thomas. 2018. "The Gravity Equation in International Trade: An Explanation." Journal of Political Economy 126 (1): 150–77.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong et al. 2022a. "Social Capital I: Measurement and Associations with Economic Mobility." *Nature* 608 (7921): 108–21.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong et al. 2022b. "Social Capital II: Determinants of Economic Connectedness." *Nature* 608 (7921): 122–34.

Chiovelli, Giorgio, Stelios Michalopoulos, and Elias Papaioannou. 2018. "Landmines and Spatial Development." NBER Working Paper 24758.

Christakis, Nicholas A., and James H. Fowler. 2009. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives.* New York: Little, Brown.

Conwell, Lucas. 2022. "Are There Too Many Minibuses in Cape Town? Privatized Provision of Public Transit." Unpublished.

Costinot, Arnaud, Dave Donaldson, and Cory Smith. 2016. "Evolving Comparative Advantage and the Impact of Climate Change in Agricultural Markets: Evidence from 1.7 million Fields around the World." *Journal of Political Economy* 124 (1): 205–48.

Costinot, Arnaud, and Andrés Rodríguez-Clare. 2014. "Trade Theory with Numbers: Quantifying the Consequences of Globalization." In *Handbook of International Economics*, Vol. 4, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 197–261. Amsterdam: North-Holland.

Coury, Michael, Toru Kitagawa, Allison Shertzer, and Matthew Turner. 2022. "The Value of Piped Water and Sewers: Evidence from 19th Century Chicago." NBER Working Paper 29718.

Couture, Victor, Jonathan Dingel, Allison Green, and Jessie Handbury. 2020. "Quantifying Social Interactions Using Smartphone Data." Unpublished.

Davis, Donald R., and David E. Weinstein. 2002. "Bones, Bombs, and Break Points: The Geography of Economic Activity." *American Economic Review* 92 (5): 1269–89.

Dekle, Robert, Jonathan Eaton, and Samuel Kortum. 2008. "Global Rebalancing with Gravity: Measuring the Burden of Adjustment." *IMF Staff Papers* 55 (3): 511–40.

Desmet, Klaus, Robert E. Kopp, Scott A. Kulp, Dávid Krisztián Nagy, Michael Oppenheimer, Esteban Rossi-Hansberg, and Benjamin H. Strauss. 2018. "Evaluating the Economic Cost of Coastal Flooding." NBER Working Paper 24918.

Desmet, Klaus, Dávid Krisztián Nagy, and Esteban Rossi-Hansberg. 2018. “The Geography of Development.” Journal of Political Economy 126 (3): 903–83.

Diamond, Rebecca, and Tim McQuade. 2019. “Who Wants Affordable Housing in Their Backyard? An Equilibrium Analysis of Low-Income Property Development.” *Journal of Political Economy* 127 (3): 1063–1117.

Dijkstra, Edsger Wybe. 1959. “A Note on Two Problems in Connexion with Graphs.” *Numerische Mathematik* 1: 269–71.

Dingel, Jonathan I., and Brent Neiman. 2020. “How Many Jobs Can Be Done at Home?” *Journal of Public Economics* 189: 104235.

Dingel, Jonathan I., and Felix Tintelnot. 2020. “Spatial Economics for Granular Settings.” NBER Working Paper 27287.

Disdier, Anne-Celia, and Keith Head. 2008. “The Puzzling Persistence of the Distance Effect on Bilateral Trade.” Review of Economics and Statistics 90 (1): 37–48.

Donaldson, Dave. 2018. “Railroads of the Raj: Estimating the Impact of Transportation Infrastructure.” *American Economic Review* 108 (4–5): 899–934.

Donaldson, Dave, and Richard Hornbeck. 2016. “Railroads and American Economic Growth: A ‘Market Access’ Approach.” Quarterly Journal of Economics 131 (2): 799–858.

Donaldson, Dave, and Adam Storeygard. 2016. “The View from Above: Applications of Satellite Data in Economics.” Journal of Economic Perspectives 30 (4): 171–98.

Ducruet, César, Réka Juhász, Dávid Krisztián Nagy, and Claudia Steinwender. 2020. “All Aboard: The Effects of Port Development.” NBER Working Paper 28148.

Duranton, Gilles, and Matthew A. Turner. 2011. “The Fundamental Law of Road Congestion: Evidence from US Cities.” *American Economic Review* 101 (6): 2616–52.

Eaton, Jonathan, and Samuel Kortum. 2002. “Technology, Geography and Trade.” *Econometrica* 70 (5): 1741–79.

Eckert, Fabian, and Michael Peters. 2022. “Spatial Structural Change.” NBER Working Paper 30489.

Faber, Benjamin, and Cecile Gaubert. 2019. “Tourism and Economic Development: Evidence from Mexico’s Coastline.” *American Economic Review* 109 (6): 2245–93.

Fajgelbaum, Pablo D., and Cecile Gaubert. 2018. “Optimal Spatial Policies, Geography and Sorting.” *Quarterly Journal of Economics* 135 (2): 959–1036.

Fajgelbaum, Pablo D., Pinelopi K. Goldberg, Patrick J. Kennedy, and Amit K. Khandelwal. 2020. “The Return to Protectionism.” *Quarterly Journal of Economics* 135 (1): 1–55.

Fajgelbaum, Pablo D., and Edouard Schaal. 2020. “Optimal Transport Networks in Spatial Equilibrium.” *Econometrica* 88 (4): 1411–52.

Fan, Jingting, Yi Lu, and Wenlan Luo. 2021. “Valuing Domestic Transport Infrastructure: A View from the Route Choice of Exporters.” Review of Economics and Statistics. https://doi.org/10.1162/rest_a_01084.

Fuchs, Simon, and Woan Foong Wong. 2022. “Multimodal Transport Networks.” Federal Reserve Bank of Atlanta Working Paper 2022-13.

Fujita, Masahisa, Paul Krugman, and Anthony J. Venables. 1999. *The Spatial Economy: Cities, Regions, and International Trade.* Cambridge, MA: MIT Press.

Ganapati, Sharat, Woan Foong Wong, and Oren Ziv. 2020. “Entrepôt: Hubs, Scale, and Trade Costs.” NBER Working Paper 29015.

Gaubert, Cecile, Patrick M. Kline, and Danny Yagan. 2021. “Place-Based Redistribution.” NBER Working Paper 28337.

Glaeser, Edward L., and Joshua D. Gottlieb. 2009. “The Wealth of Cities: Agglomeration Economies and Spatial Equilibrium in the United States.” *Journal of Economic Literature* 47 (4): 983–1028.

Head, Keith, and Thierry Mayer. 2013. “Gravity Equations: Workhorse, Toolkit, and Cookbook.” In *Handbook of International Economics*, Vol. 4, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 131–95. Amsterdam: North-Holland.

Heblich, Stephan, Stephen J. Redding, and Daniel M. Sturm. 2020. “The Making of the Modern Metropolis: Evidence from London.” Quarterly Journal of Economics 135 (4): 2059– 2133.

Heiland, Inga, Andreas Moxnes, Karen Helene Ulltveit-Moe, and Yuan Zi. 2019. “Trade from Space: Shipping Networks and the Global Implications of Local Shocks.” CEPR Discussion Paper 14193.

Helpman, Elhanan. 1998. “The Size of Regions.” In *Topics in Public Economics: Theoretical and Applied Analysis*, edited by David Pines, Efraim Sadka, Itzhak Zilcha, 33–54. Cambridge, UK: Cambridge University Press.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102 (2): 994–1028.

Jaworski, Taylor, and Carl T. Kitchens. 2019. "National Policy for Regional Development: Historical Evidence from Appalachian Highways." Review of Economics and Statistics 101 (5): 777–90.

Kleinman, Benny, Ernest Liu, and Stephen J. Redding. 2021. "Dynamic Spatial General Equilibrium." NBER Working Paper 29101.

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.

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

Lee, Sun Kyoung. 2022. "When Cities Grow: Urban Planning and Segregation in the Prewar US." Unpublished.

Manson, Steven. 2020. "IPUMS National Historical Geographic Information System: Version 15.0." IPMUS. http://doi.org/10.18128/D050.V15.0 (accessed March 1, 2023).

McCallum, John. 1995. "National Borders Matter: Canada-U.S. Regional Trade Patterns." American Economic Review 85 (3): 615–23.

Monte, Ferdinando, Stephen J. Redding, and Esteban Rossi-Hansberg. 2018. "Commuting, Migration, and Local Employment Elasticities." *American Economic Review* 108 (12): 3855–90.

Nordhaus, William, and Xi Chen. 2006. "Geographically Based Economic Data (G-Econ)." Yale University. https://gecon.yale.edu/ (accessed March 1, 2023).

Peters, Michael. 2022. "Market Size and Spatial Growth—Evidence from Germany's Post-war Population Expulsions." *Econometrica* 90 (5): 2357–96.

Redding, Stephen J. 2016. "Goods Trade, Factor Mobility and Welfare." *Journal of International Economics* 101: 148–67.

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

Redding, Stephen, and Anthony J. Venables. 2004. "Economic Geography and International Inequality." Journal of International Economics 62 (1): 53–82.

Roback, Jennifer. 1982. "Wages, Rents, and the Quality of Life." Journal of Political Economy 90 (6): 1257–78.

Rosen, Sherwin. 1979. "Wage-Based Indexes of Urban Quality of Life." *Current Issues in Urban Economics*: 74–104.

Sethian, J. A. 1999. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Vol. 3. Cambridge, UK: Cambridge University Press.

Severen, Christopher. 2021. "Commuting, Labor, and Housing Market Effects of Mass Transportation: Welfare and Identification." Review of Economics and Statistics. https://doi.org/10.1162/rest_a_01100.

Steinwender, Claudia. 2018. "Real Effects of Information Frictions: When the States and the Kingdom Became United." *American Economic Review* 108 (3): 657–96.

Tombe, Trevor, and Xiaodong Zhu. 2019. "Trade, Migration, and Productivity: A Quantitative Analysis of China." *American Economic Review* 109 (5): 1843–72.

Topalova, Petia. 2010. "Factor Immobility and Regional Impacts of Trade Liberalization: Evidence on Poverty from India." American Economic Journal: Applied Economics 2 (4): 1–41.

Tsitsiklis, J. N. 1995. "Efficient Algorithms for Globally Optimal Trajectories." IEEE Transactions on Automatic Control 40 (9): 1528–38.

US Department of Transportation, Bureau of Transportation Statistics; and, US Department of Commerce, US Census Bureau. (2020-08). 2017 Commodity Flow Survey Datasets 2017 CFS Public Use File (PUF). Available at: https://data.nber.org/transportation/CommoditiesData/States/states_bilateral_trade2017.dta.

Zárate, Román D. 2022. "Spatial Misallocation, Informality, and Transit Improvements: Evidence from Mexico City." World Bank Policy Research Working Paper 9990.

Journal of Economic Perspectives—Volume 37, Number 2—Spring 2023—Pages 29–52

Neighborhood Change, Gentrification, and the Urbanization of College Graduates

Victor Couture and Jessie Handbury

There has been a striking reversal in where college graduates choose to live within the largest US cities. For most of the twentieth century, Americans who could afford it moved to the suburbs. At some point after 1980, college graduates started moving back downtown. This urban revival intensified at the turn of the twenty-first century, even as the suburbanization of the US population as a whole continued unabated. Accelerating inflows of college graduates transformed downtown neighborhoods in almost all large US cities, raising policy concerns over housing affordability in gentrifying areas. The share of downtown residents with a college degree rose threefold between 1980 and 2017, from 15 to 45 percent, and downtown areas reverted from being the least-educated to being the most-educated areas of US cities. This gentrification of the United States's downtown areas had a strong age and racial bias. Over the last few decades, college graduates who are young and white experienced much larger changes in their propensity to live downtown than any other demographic group. Now, in the post-pandemic era, there are early signs of renewed suburban attractiveness and there may be yet another reversal in college graduates' location choices. In this article, we discuss these changes in where the college-educated choose to live within cities and their consequences.

A number of factors can explain these reversals in city centers' fortune over the last century. We focus on the role of transportation technology and income

■ *Victor Couture is Assistant Professor of Economics, University of British Columbia, Vancouver, Canada. Jessie Handbury is Associate Professor in Real Estate, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania. Their email addresses are victor.couture@ubc.ca and handbury@wharton.upenn.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.29.

growth, but we also mention other forces which may have played a role. Why were college graduates suburbanized in 1980, prior to urban revival? In the early-to-mid–twentieth century, the mass production of cars made the suburbs attractive places for Americans rich enough to afford this new transportation technology (Glaeser, Kahn, and Rappaport 2008). Cars allow for lower density living by removing the need for long walks to and from transit stops. Car-owning households could move to low-density residential suburbs, enjoy larger houses, and keep their commute time short by driving to work.

In the late twentieth century, rising income inequality and delayed family formation contributed to reversing these trends and re-urbanizing rich and educated households (Couture and Handbury 2020). As the rich got richer, their time became more valuable. They also had more disposable income to spend on services like restaurants, bars, gyms, and beauty salons. They sought to avoid spending valuable time commuting and found proximity to downtown concentration of amenities and jobs more attractive. Downtown amenities are also meeting places for networking, friendships, and dating, which makes them particularly popular with richer people who are unmarried and childless. As a result, the rising incomes of young college graduates and their reduced propensity to marry and have children early in life both contributed to downtown gentrification.

The gentrification of US downtowns had important welfare consequences. When higher-income people move to a neighborhood, its local amenities adjust to match their tastes and budgets. Improving schools and local services make those neighborhoods even more attractive to higher-income people, thus amplifying neighborhood change. Because the rich can outbid the poor for housing in neighborhoods that become more desirable, and because new housing is often difficult to build in land-constrained downtowns, rapidly rising housing costs further exacerbate the welfare impact of downtown gentrification.[1] Indeed, rising house prices are a controversial by-product of neighborhood gentrification. Cities across the country responded by implementing a variety of policies to maintain housing affordability in gentrifying neighborhoods, from rent control to various incentives to build affordable housing. These antigentrification policies have generally seen only modest success.

Rising housing costs almost certainly hurt the poor more than the rich. In that sense, gentrification may reinforce welfare inequality (Couture et al. 2019). Spatial sorting may also affect future inequality, because one's residential neighborhood growing up has a causal effect on later-life outcomes ranging from income to crime (in this journal, Chyn and Katz 2021). When the highest-income households live in the highest opportunity areas of cities, children of those households are more likely to remain high-income in adulthood and poorer children are less likely to come

[1] A literature also documents differences in the preferences of rich and poor households and in how they value urban amenities. High-income individuals incur higher time-cost of travel (Small and Verhoef 2007) and they have stronger preferences for neighborhood characteristics like school quality (Bayer, Ferreira, and McMillan 2007), crime rates (Ellen, Horn, and Reed 2019), access to jobs (Su 2022), and access to nontradable services like restaurants, gyms, and personal services (Couture and Handbury 2020). Preferences for neighborhood characteristics vary by race as well (Baum-Snow and Hartley 2020).

out of poverty. The reduction in intergenerational mobility due to spatial sorting has likely contributed to the dramatic growth in income inequality over the past 50 years (Fogli and Guerrieri 2019; for a survey of the literature on between-city sorting and inequality, see Diamond and Gaubert 2022).

Could the urban revival still generate beneficial social connections between college-educated newcomers and existing residents? We find that less-educated people who remained in gentrifying downtowns did experience higher exposure to college graduates within their neighborhoods. In general, however, college graduates clustered into select downtown neighborhoods instead of mixing with less-educated incumbents. As a result, neighborhood-level sorting across educational lines sharpened between 1980 and 2017. This sorting behavior likely limits social interactions between the young, white, college-educated gentrifiers and the less educated, often minority incumbents.

Changes in Socioeconomic Spatial Sorting: 1980–2017

Who Lives Near the City Center?

We begin by characterizing the location decisions of college graduates in US cities since 1980. We define cities using Core-Based Statistical Areas or “CBSAs.” CBSAs are geographic areas designed by the US Office of Management and Budget to contain sets of contiguous counties tied to urban centers by commuting. We do not know exactly where each household resides in each CBSA, but the Census Bureau provides data on the number of households and individuals—both in aggregate and by education level, age, and race—that reside in small areas called Census tracts. Census tracts are drawn so as to contain around 4,000 households. We sometimes refer to Census tracts as “neighborhoods.”[2]

We are interested in how neighborhoods changed differently depending on how far they are from the city center. To study this, we first fix the center of each city at the coordinates of its city hall (as defined in Holian 2019). We then normalize the distance to these centers to a population-weighted metric that ranges between zero and one in all of our sample cities. This distance metric is equal to the share of the population for the city as a whole that resided in Census tracts whose centroids are at the same distance as or closer to the city center than that tract in 2000. The tract whose centroid is furthest from the city center in a given Core-Based Statistical Areas has a distance of one, while the tract whose centroid is closest to the city center has a distance close to zero (equal to that tract’s own share of the 2000 CBSA population).

Our analysis focuses on the 100 Core-Based Statistical Areas with the highest populations in 2000. Between 1980 and 2017, the population of these large US cities grew by 55 percent. This growth was not accompanied by a proportionate

[2] The boundaries of Core-Based Statistical Areas and Census tracts shift over time. To avoid the contamination of our results by changes in CBSA and tract boundaries, we fix the boundaries of CBSAs and tracts to their 2010 definitions.

densification of city centers, but instead occurred through sprawl. While neighborhoods near city centers grew by less than 10 percent on average between 1980 and 2017, the population in the outer suburbs grew by over 50 percent.[3]

The location choices of college graduates, however, differed from those of the general population. While aggregate population growth was monotonically higher in neighborhoods further from city centers, college-educated population growth was bifurcated between the city center and outer suburbs. Between 1980 and 2017, the number of college graduates grew threefold at the city center and over fourfold in the outer suburbs, but less so at intermediate distances from the city center. The contrasting patterns of aggregate and college-educated population growth resulted in varied changes in the mix of residents in urban cores versus the suburbs. In particular, the growth in the college-educated population in central neighborhoods in the backdrop of limited population growth overall resulted in dramatic increases in the share of downtown residents with a college degree.

Panel A of Figure 1 shows how the share of college-educated residents in neighborhoods at different distances from the city center changed between 1980 and 2017, revealing three distinct patterns. First, the average college share rose steadily from 1980 to 2017 at all distances from the city center. This pattern reflects both that the US population became more educated over this time period and that college graduates became increasingly concentrated in large cities (Moretti 2012).

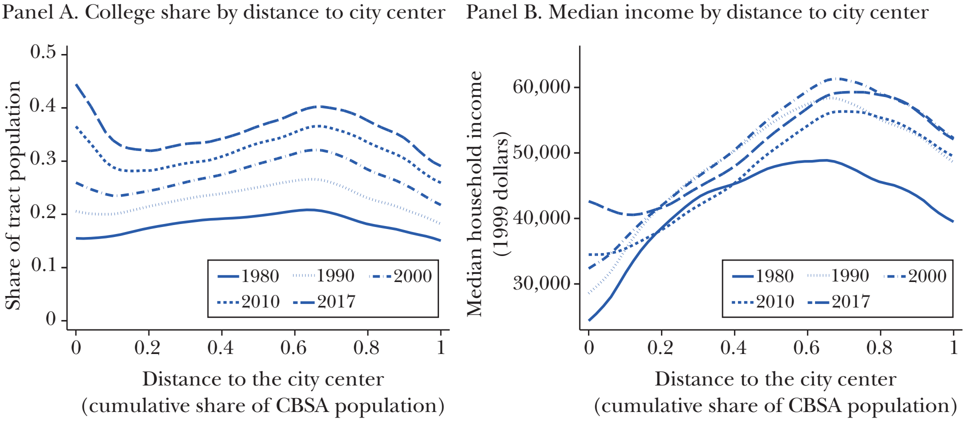
Second, the share of college graduates living near city centers increased sharply. In 1980, the innermost neighborhoods had the lowest college share. From 1990 to 2000, there was a small uptick in the share of college graduates living downtown, which accelerated rapidly from 2000 to 2010. By 2017, the initial 1980 sorting patterns had entirely reversed and the innermost neighborhoods had the highest college-educated shares.

Finally, there was a sharpening of spatial sorting by education from 1980 to 2017. In 1980, the college-share gradient by distance to city centers is almost flat. By 2017, the variation in college shares by distance to city centers had risen dramatically, with a gap of almost ten percentage points between the highest college shares near city centers, and the lowest college shares in the inner suburbs.

We use college attainment as our primary measure of resident socioeconomic status, but panel B of Figure 1 shows qualitatively similar patterns in neighborhood resident income. There was a substantial uptick in median income of neighborhoods near city centers, especially in the later period from 2010 to 2017. The uptick of central-city resident income, however, was smaller than that for education, and the highest income neighborhoods are still in the suburbs as of 2017. One possible explanation for this difference in sorting by education and income is that the college-educated moving downtown tend to be young (as we will discuss in the next subsection), and therefore earn relatively lower incomes.

[3] We refer interested readers to our online Appendix for more detail on these aggregate population trends, as well as additional results showing the robustness of our empirical patterns.

Figure 1
Neighborhood Socioeconomic Status by Distance to City Center



Source: NHGIS Census (1980, 1990, 2000) & American Community Survey (2008–2012, 2015–2019) (Manson et al. 2022); Longitudinal Tract Data Base (Logan, Xu, and Stults 2014; Holian 2019).
Note: This figure plots the socioeconomic status of Census tract residents by distance to the city center in each decade from 1980 to 2017. Panel A shows the college-educated share of the Census tract population by distance to the city center in each decade from 1980 to 2017. Panel B shows the real median household income by distance to the city center. Each line is a nonparametric kernel regression of Census tract-level demographic data from the largest 100 cities (the Core-Based Statistical Areas [CBSAs] with the highest populations in 2000). Each kernel regression observation is weighted by tract population. Distance is measured as the share of the city residents that live at least as close to the city center, which is zero at the center and one at the furthest point in the metropolitan area.

Who Is Moving Downtown?

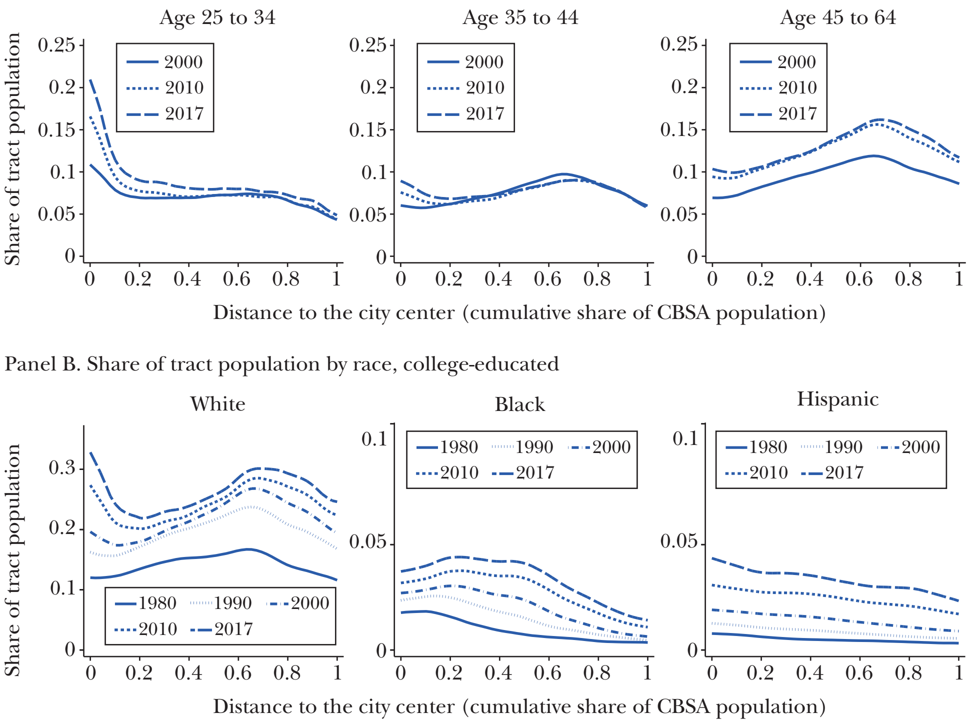
We now break down the college-educated into finer demographic groups and identify a significant age and racial bias in the United States's urban revival (Baum-Snow and Hartley 2020; Couture and Handbury 2020). Figure 2 breaks down the college-educated tract population shares depicted in panel A of Figure 1 into age and racial groups. Panel A of Figure 2 plots the population of college graduates in each age group as a share of the overall Census tract population against the population-weighted distance of the neighborhood from the city center in 2000, 2010, and 2017 (2000 is the earliest year for which tract-level population data by age and education is available). The youngest cohort of college graduates, aged 25 to 34, accounts for the vast majority of college share growth near city centers between 2000 and 2017. For the oldest college-educated cohort, aged 45 to 64, we see almost no change in sorting patterns between 2000 to 2017.

Panel B of Figure 2 plots the population of white, Black, and Hispanic college graduates as a share of the overall Census tract population against distance to the city center in each decade from 1980 to 2017. Here, the uptick in college graduate shares downtown is overwhelmingly driven by white college graduates, with smaller increase for Hispanics. Notably, Black college graduates display the opposite sorting

Figure 2
College Share by Distance to City Center, Demographic Breakdowns

Panel A. Share of tract population by age group, college-educated

Panel B. Share of tract population by race, college-educated



Source: NHGIS Census (1980, 1990, 2000) & American Community Survey (2008–2012, 2015–2019) (Manson et al. 2022); Longitudinal Tract Data Base (Logan, Xu, and Stults 2014); Holian (2019).
Note: This figure plots the share of the Census tract population in given demographic groups by distance to the city center. Panel A displays what fraction of the Census tract population is both college-educated and within the specified age group. Panel B displays what fraction of the Census tract population is both college-educated and the specified race or ethnicity. The categories are not exclusive: both non-Hispanic White and Hispanic White residents are classified as White, and Hispanic residents of any race are classified as Hispanic. Each line is a nonparametric kernel regression of Census tract-level demographic data from the largest 100 cities, defined as the Core-Based Statistical Areas (CBSAs) with the highest populations in 2000. Each kernel regression observation is weighted by tract population. Distance is measured as the share of the city residents that live at least as close to the city center, which is zero at the center and one at the furthest point in the metropolitan area.

patterns. While white college graduates are making up increasing shares of neighborhoods close to city centers and in the far suburbs, Black college graduates make up an increasing share of near-suburban neighborhoods.[4]

[4] We note that even though white college graduates saw the strongest change in sorting patterns, the downtowns of large cities became less white from 2000 to 2017, in line with general trends in the US population.

Where Are College Graduates Moving Downtown?

In general, the urban revival that we document here is a "big city" phenomenon, but it is not unique to a few large cities. Reproducing panel A of Figure 1 separately for each of the twelve largest US cities, we find a substantial uptick in the share of college graduates living near city centers in all but one city. New York, Chicago, Washington DC, and Houston experience particularly sharp rises in the downtown college share. Los Angeles, a city noted for having many centers (Redfearn 2007), is the one exception, but even there the neighborhoods near city hall (the definition of the city center that we use) saw a rising share of young college graduates.

Within the central areas of big cities, urban revival was not evenly distributed. Figure 3 maps this process in two individual cities, Chicago and Philadelphia. The figure shows how the college-educated population was sorted across Census tracts in each city in 1980 and 2017. The shading of each tract reflects the college-educated share of its residents in that year. Tracts with the highest college shares are in darker blue, and tracts with the lowest college shares are in pale yellow. The solid black boundaries indicate each city's downtown area, consisting of the tracts closest to the city center containing 10 percent of the city's population in 2000.

Within downtowns, we see clusters of neighborhoods with relatively high college shares (shaded in blue) in both 1980 and 2017. We also see clusters of highly educated neighborhoods in the suburbs of each city. In between these blue central and peripheral regions, there is a pale "donut" of neighborhoods with lower college graduate shares. These donut-like patterns became more pronounced from 1980 to 2017.

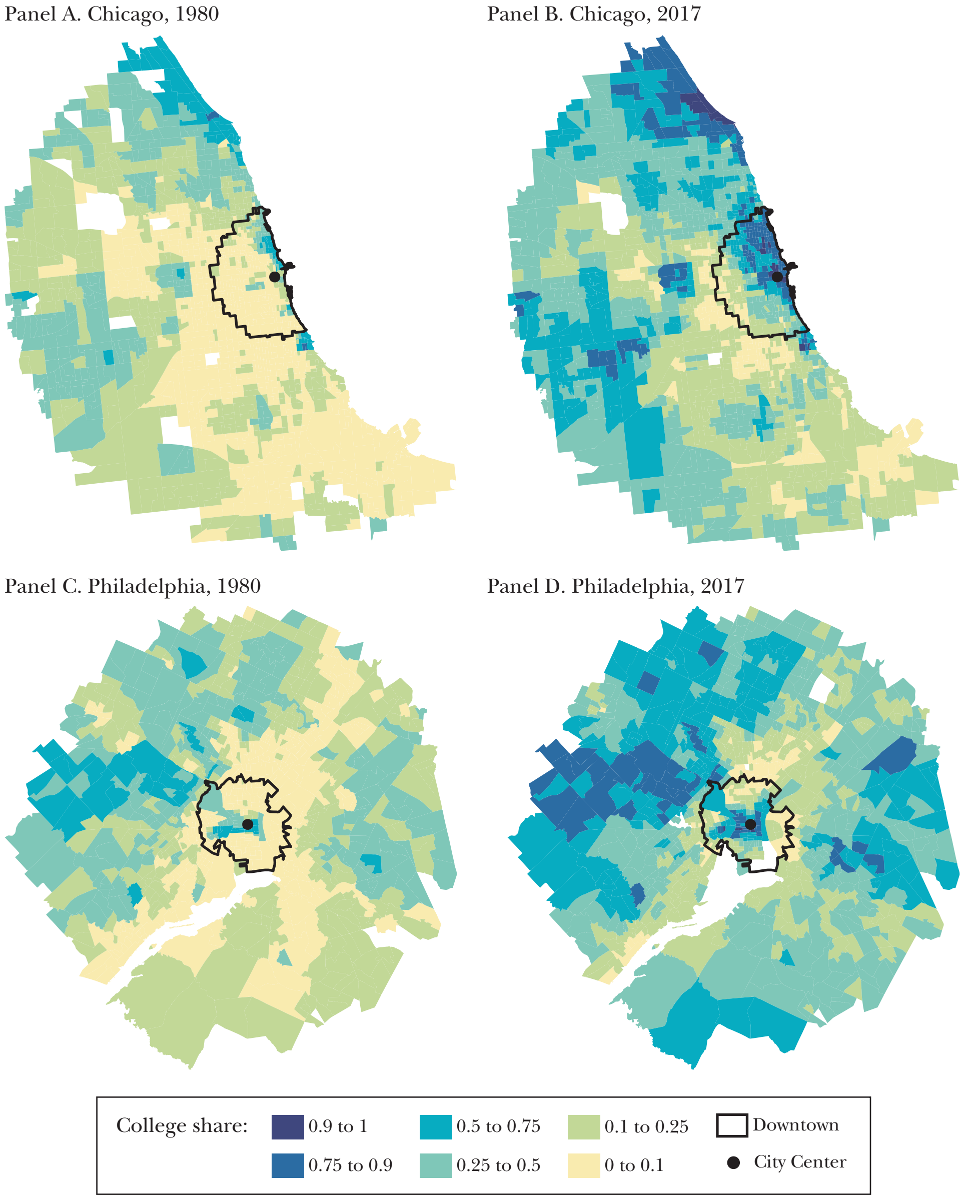
Figure 3 also shows that the growth in college share from 1980 and 2017 was not evenly distributed across neighborhoods at the same distance from the city center. Between 1980 and 2017, the city-wide college attainment rate grew from 0.17 to 0.39 in Chicago and 0.16 to 0.36 in Philadelphia. A select group of downtown neighborhoods saw especially large increases in their college share relative to the growing means. These "gentrifying" neighborhoods were often adjacent to existing 1980 enclaves of college graduates (Guerrieri, Hartley, and Hurst 2013). Neighborhoods with initially low college-educated shares fell even further behind the city average college share between 1980 and 2017.

The uneven distribution of growth in educational attainment rates across neighborhoods increased the variance in college shares over time. Qualitatively, this shift is reflected in the sharper color contrast between the light and dark-shaded neighborhoods in the maps over time. Quantitatively, the gap in the college share between the 75th percentile and 25th percentile of Census tracts increased from 0.18 to 0.34 in Chicago and from 0.19 to 0.33 in Philadelphia from 1980 to 2017. These interquartile differences are representative of that observed overall in the largest 100 Core-Based Statistical Areas; in this group, the interquartile range increased from 0.17 to 0.31.

In summary, while the general US population has continued to suburbanize since 1980, college graduates bucked this trend and displayed rising propensity to live near city centers from the 1990s onward. This urban revival reversed decades of suburbanization of the college-educated and represents a sharpening in sorting by

Figure 3

Variation in College Share across Census Tracts



Panel A. Chicago, 1980

Panel B. Chicago, 2017

Panel C. Philadelphia, 1980

Panel D. Philadelphia, 2017

Source: NHGIS Census (1980, 1990, 2000) & American Community Survey (2008–2012, 2015–2019) (Manson et al. 2022); Longitudinal Tract Data Base (Logan, Xu, and Stults 2014); Holian (2019).

Note: This figure maps the share of Census tract residents who are college educated relative to the Core-Based Statistical Area average. The black dot shows the city center. The solid black boundaries delineate each city's downtown area, consisting of the tracts closest to the city center containing 10 percent of the city's population in 2000. The maps show the tracts in Chicago and Philadelphia comprising the 60 percent of each city's population that live closest to the city center. The tract college-educated shares are demeaned using the share of the CBSA population that is college-educated.

education within cities, driven almost entirely by younger cohort of white college graduates. We now turn to investigating the forces that may explain these changes in sorting patterns.

Forces behind Spatial Sorting

How do rich and poor households choose where to live? We focus on how changes in transportation technology (LeRoy and Sonstelie 1983; Glaeser, Kahn, and Rappaport 2008) and changes in the income distribution (Couture et al. 2019) can explain the evolution of spatial sorting patterns. The introductions of transportation technologies like trains, cars, and bicycles are a natural candidate to explain changes in household location choices. Indeed, cities primarily exist to reduce physical distance between people and to facilitate social interactions. Glaeser (2020) and Heblich, Redding, and Sturm (2020) show how innovations in transportation, from horse-drawn omnibus to steam- and electric-powered urban railways, shaped the modern city in the nineteenth century. These early public transit systems, by allowing individuals to live farther from where they worked, saw the emergence of the familiar city structure in which an urban core with high job density is surrounded by residential suburbs. Rising income inequality is also a natural candidate to explain changes in spatial sorting. We expect travel costs and travel preferences (that is, the type of venues visited) to change as households get richer or poorer.

To illustrate the key determinants of sorting within cities, we present a basic model of a city in which travel costs determine the location choices of different income groups. We also discuss existing empirical evidence in support of the model's prediction.

Baseline: Monocentric City Model

We start by establishing baseline theoretical sorting patterns within a monocentric and linear city. In this setting, individuals belonging to different income groups choose where to live. We assume that each individual consumes one unit of housing, so our informal discussion of this model ignores issues of housing size and quality. We assume that all jobs and amenities are located at the city center, so the costs of commuting to work and of traveling to consume amenities rise with distance to the city center (as in Brueckner, Thisse, and Zenou 1999). In reality, of course, not all jobs and amenities are located downtown, but the density of jobs and nontradable service amenities is highest near city centers. Moreover, urban amenities, specifically nontradable services like restaurants, bars, gyms, and personal services, are the key factor attracting young college graduates towards city centers (Couture and Handbury 2020).

We assume that each individual commutes once each day to the city center to work and, depending on their travel preferences, makes additional city center trips to consume amenities. In the data, specifically the 2009 National Household Transportation Survey, working-age residents of large cities take about as many trips to nontradable services as they take work-related trips, so both types of trips are likely important in thinking about location decisions.

The model has three income groups: high-, middle-, and low-income. Travel preferences vary by income in two different ways. First, we assume that higher income individuals have a stronger taste for city center amenities, so the number of amenity trips rises with income. In the National Household Transportation Survey data, high-income workers take almost twice as many trips to nontradable services as low-income workers. Second, we assume that higher income individuals have higher value of time, so the opportunity cost of commuting to work or of traveling to consume amenities rises with income. A large empirical literature, reviewed in (Small and Verhoef 2007), verifies this assumption and finds that value of travel time rises roughly proportionally with wages.

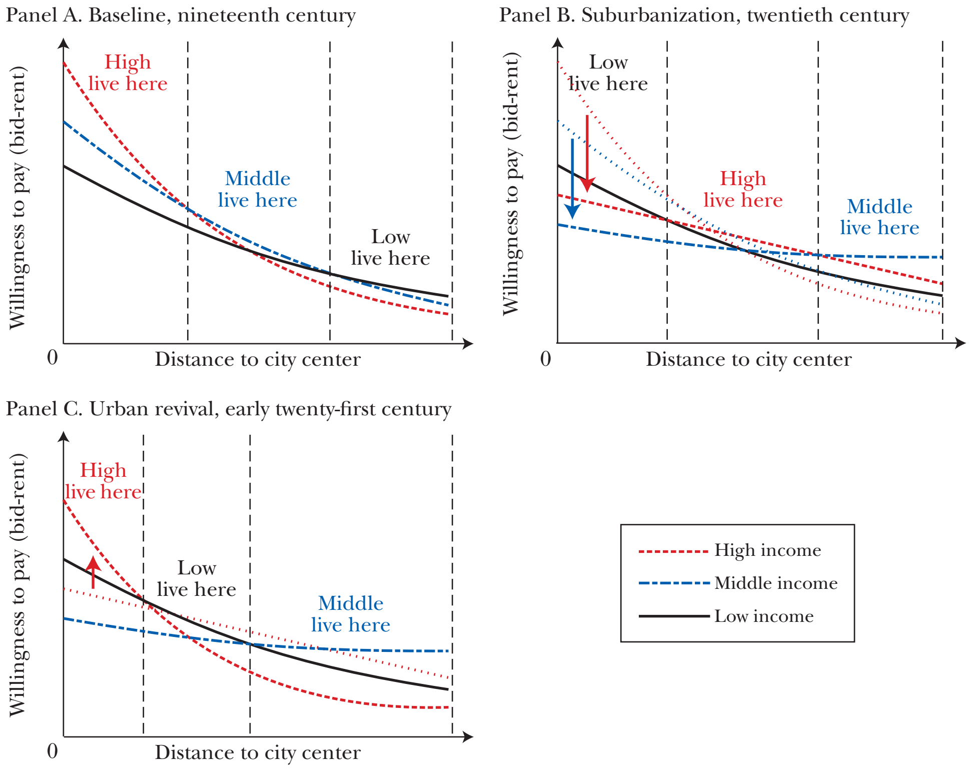
These assumptions about travel costs and preferences suggest that willingness to pay to live at different distances from the city center will vary by income. In turn, differences in willingness to pay determine sorting patterns—and in particular, which income group lives closer to the city center. In the canonical monocentric city model, willingness to pay, as a function of distance to city center, is called the "bid-rent" function. This terminology illustrates how different groups "bid" for housing at different locations in the city. In equilibrium, each location in the city is inhabited by the income group with the highest bid-rent function, or willingness to pay, at that location.

Panel A of Figure 4 depicts the bid-rent function of each income group in the baseline model described above. The horizontal axis is distance from the city center, and the vertical axis is willingness to pay ("bid-rent function") to live at each location. Given our assumption that all jobs and amenities are at the city center, the highest willingness to pay is for housing at the city center, where travel costs are lowest. The bid-rent function then declines with distance to the city center in exact proportion to the increase in travel costs from longer work and amenity trips. The slope of the bid-rent function is the negative of travel costs per unit distance, multiplied by the total number of trips to the city center (work plus amenity trips). A steeper bid-rent function means a higher willingness to pay to live closer to city centers. In panel A, the bid-rent function is steepest for high-income individuals. That is, higher-income people have the highest willingness to pay to live near city centers because they take more city center trips to consume amenities than lower-income people and incur a higher cost of travel time.

Each location is inhabited by the income group with the highest willingness to pay to live there, because that group will outbid other groups for housing in that location. High-income individuals live closest to the city center: their bid-rent curve is higher than that of other income groups near city centers, so they are willing to pay the most to live there. Low-income individuals live furthest from city centers, in the far suburbs, where their bid-rent curve is higher than that of any other groups. Middle-income individuals live in between. Our model's baseline sorting patterns are therefore strictly monotonic by income. These monotonic sorting patterns are observed in many European cities (Brueckner, Thisse, and Zenou 1999), in the United States before the suburbanization era (for late nineteenth century income

Figure 4

Sorting Patterns by Income in the Monocentric City Model



Source: Authors' illustration.

Note: The figure displays the bid rent functions (willingness to pay) of high-income, middle-income, and low-income households to live at different distances from the city center at location zero, under three scenarios described in this section. Panel A displays the baseline monotonic sorting pattern. Panel B shows the flattening of high- and middle-income residents' bid-rent function following the introduction of the car in the early to mid-twentieth century. Panel C shows the steepening of high income residents' bid-rent function as a result of top income growth in the late twentieth century that made them richer.

gradients from city centers in large US cities, see Lee and Lin 2018), and possibly also in ancient settings (Gupta and Halket 2021).

These baseline sorting patterns, where higher-income people have a higher valuation for proximity to city center jobs and amenities, depend on our assumptions that richer people have both a higher opportunity cost of travel time and a higher propensity to consume amenities. If we modify these assumptions to better match reality at a given point in time—for instance, by assuming that richer people can afford cars that lower their cost of travel time—then the model can deliver different sorting patterns. We can thus use this model to study what factors might generate the deviations from monotonic income sorting observed in the data shown in panel A of Figure 1.

Suburbanization: Improvements in Transportation Technology

We first discuss the origins of the sorting patterns in place in 1980, prior to urban revival. In particular, we will argue that changes in transportation technology contributed to the urbanization of the poor and the suburbanization of the rich in the twentieth century. We focus on the impact of mass production of fast private motorized vehicles in the early to mid-twentieth century. The car gave people the ability to live further from work, in lower-density neighborhoods, unconstrained by the need to walk to and from transit stops. But into the middle of the twentieth century, car transportation technology was often only affordable to high- and middle-income households. Low-income households still relied on slower public transit networks. These transit networks were more developed near city centers, which had sufficient population density to support transit stops within walking distance of jobs and residences (LeRoy and Sonstelie 1983; Glaeser, Kahn, and Rappaport 2008).

In the context of our model, lower travel costs for car-owning middle- and high-income individuals make their bid-rent functions less steep relative to that of low-income households. That is, living near city center jobs and amenities becomes less valuable for car-owning households, because their travel costs are lower. This change is depicted in the flattening of the high- and middle-income bid-rent curves in panel B of Figure 4. After this change, it is low-income households, who still commute by slower public transit, who are willing to bid the most for housing near the city center. High- and middle-income households, who commute by cars, are willing to bid the most for housing in the near and far suburbs, respectively. Accordingly, low-income households move closest to the city center; there is an urbanization of poverty. Middle-income households now reside furthest from city centers, and high-income households live in between. Of course, our basic model does not allow for more than one income group to live in the same area. However, Couture et al. (2019) document U-shaped urbanization by income in every decade since 1970 (data from 1950 is incomplete), in which both high- and low-income households have a higher propensity to live downtown than middle-income households.

These sorting patterns are consistent with the data in panel A of Figure 1, which shows that in 1980, areas with the highest share of college graduates were near the midpoint of the city. Similarly, panel B of Figure 1 shows that the median household income was lowest near city centers in 1980 and highest at about 65 percent of the total distance from city centers.

Of course, the suburbanization of the United States had many causes beyond the arrival of the private car. A comprehensive review is beyond the scope of our paper, but we wanted to mention a few factors here: expanded transportation infrastructure, white flight, and income growth.[5] First, at the turn of the twentieth century, the

[5] Our present article focuses on sorting by education and income, but a larger literature studies the suburbanization of the general US population in the twentieth century. See Jackson (1987) for an

earliest US suburbs were based on streetcars, predating widespread car adoption. By mid-century, the construction of the interstate highway system further contributed to suburbanizing residential locations (Baum-Snow 2007), while the advent of trucking further suburbanized manufacturing employment (Mieszkowski and Mills 1993). Second, white flight was another notable suburbanization force during the mid-twentieth century. Boustan (2010) shows that from 1940 to 1970, white households moved out of central cities, in part as a response to the in-migration of Black households, who were moving out of the rural South and into urban areas during the second wave of the Great Migration. Finally, Margo (1992) shows that income growth contributed to post–World War II suburbanization. Similarly, Couture et al. (2019) offer a theory of how broad-based income growth from 1950 to 1980—in the middle as opposed to the top of the income distribution—allowed a larger share of households to afford spacious suburban housing.

Many other mechanisms not in our model contribute to, reinforce, and amplify these sorting patterns. We discuss some these other forces, like locally-funded schools and crime, later in the article as part of our discussion of the consequences of spatial sorting.

Urban Revival: Growth of Incomes at the Top and Other Forces

We now explain the urbanization of college-educated and high-income households in the late twentieth and early twenty-first century, focusing on the role of changes in the income distribution. Looking across US cities, Couture et al. (2019) find a strong relationship between cities that experienced more growth of top incomes and cities in which those with high incomes moved downtown between 1990 and 2014. As those with higher incomes got relatively better-off from the 1980s onward, their time became more valuable. They also had more disposable income to spend on local service amenities. So as income inequality rose, high-income households placed a higher value on being within a short travel time from downtown amenities and jobs. The existing empirical literature is largely consistent with the idea that newly-rich households moved to dense downtowns to save on travel costs. Baum-Snow and Hartley (2020) and Couture and Handbury (2020) find that the rising valuation of downtown amenity density was the key driver of the urbanization of college-educated and high-income households in the early twenty-first century. Su (2022) finds instead that the rising value of job density (shorter commutes) was more important.

In the context of our basic model, higher growth of the top incomes raises both the number of amenity trips and the travel costs per unit distance of high-income households. Both forces raise the valuation of high-income households for locations close to city centers where jobs and amenities are located. If the bid-rent function of high-income individuals becomes steep enough, they will live closest to the city center to save on travel costs, as shown in panel C of Figure 4. Middle-income

authoritative history of suburbanization in the United States, and in this journal, Mieszkowski and Mills (1993) for a review of the literature on the causes of suburbanization.

individuals still live furthest from to the center, with low-income people living in between.

These sorting patterns are consistent with the facts documented earlier. Panels A and B of Figure 1 respectively showed rising shares of college graduates and household incomes near city centers from 1990 to 2017. By 2017, the highest share of college graduates was near city centers, and the lowest share was just outside of downtowns. Education gradients vary by city, but many individual cities feature the "donut"-like sorting patterns derived from our model above.[6] In those cities, the highest-income households live downtown, the lowest-income households live in the inner suburbs, and middle-income households live in the outer suburbs. The sorting by income is not as pronounced as it is for education, likely because college graduates moving downtown tend to be young, and younger people have lower incomes. Taking again Chicago and Philadelphia as examples, Figure 3 showed the emergence of a dark blue area of high college shares around city centers from 1980 to 2017, surrounded by some of the least educated neighborhood in these cities, in lighter yellow shades.

Naturally, additional forces also contributed to the urbanization of college graduates in the late twentieth and early twenty-first century, notably delayed family formation (Couture and Handbury 2020; Moreno-Maldonado and Santamaria 2021) and declining urban crime (Ellen, Horn, and Reed 2019). Young college graduates, who urbanized most rapidly during this time period, were ten percentage points more likely to be solo (unmarried and childless) in 2014 than in 1990 (Couture and Handbury 2020). Solo individuals are by far the most urbanized family type and they have the highest propensity to travel to nontradable service amenities like restaurants and bars that are meeting places for social connections, friendships, and dating. They also have less need for the additional housing space and generally better schools available to suburban residents. In other words, the rising share of young college graduates who delay family formation to live solo is consistent with the rising urbanization rate of that group. Another plausible hypothesis is "reverse" white flight—the idea that younger cohorts of white college graduates have a greater taste for residing in minority neighborhoods (prevalent in downtowns) than older cohorts—although we are not aware of formal tests of this hypothesis. Finally, Ellen, Horn, and Reed (2019) show that richer and more educated individuals also have higher valuations for low crime environments. As a result, declining central city crime from the 1990s onwards is consistent with the urbanization of the rich and college-educated. Later in this paper, we discuss changes in crime rates as a consequence—rather than a cause—of within-city sorting.

[6] Figure A.2 shows education gradients for the twelve largest cities. Glaeser, Kahn, and Rappaport (2008) discuss the historical origin of differences in sorting patterns. For instance, New York City was built around public transit, so unlike Los Angeles, which was built around the car, New York City always retained a sizable share of richer households near its center in Manhattan.

Consequences of Spatial Sorting

The mix of people who reside in a location determines who can access the amenities, like schools and local services, available in that location. This residential mix also affects what those amenities are and how much this access costs (in part, how much it costs to live in that location). These changes in amenities and house prices resulting from within-city sorting affect the welfare of high- and low-income households differently. The mix of residents in a neighborhood is also a key determinant of social mobility—that is, of the likelihood that poor children from that neighborhood rise out of poverty. In this section, we discuss these effects and then offer new evidence on whether the urbanization of college graduates was associated with greater residential mixing, and less segregation, by education.

Endogenous Amenities

As the demographic mix of residents in a neighborhood changes, so too do the amenities available there. Amenities catering to local tastes then attract further in-migration of residents with similar tastes, thereby amplifying spatial sorting patterns. For instance, the suburbanization of high-income households in the mid-twentieth century likely led to better public schools there which, in turn, attracted further suburbanization of households that placed a high value on education (Mieszkowski and Mills 1993). More generally, local financing of public goods may anchor richer households to suburban municipalities.

Privately-offered nontradable amenities, such as restaurants, bars, gyms, private schools, and beauty salons, have played a more central role in explaining the urbanization of the high-income and well-educated. These amenities have scale economies, so their availability correlates with residential population density. Downtowns offer foot traffic from both high residential density and daytime office workers, so they feature the densest, most diverse, and highest-quality mix of these nontradable service amenities. College graduates, particularly the young college graduates whose shifting location choices have gentrified downtowns, spend more at these venues and take more trips to them than other groups. The bias of the young and college-educated towards these amenities has increased over time as this group earns higher incomes and delays marriage and childbearing (Couture and Handbury 2020).

With venues serving as meeting places, assortative matching in dating and friendship act as another mechanism endogenously reinforcing the urbanization of the rich and college-educated (Moreno-Maldonado and Santamaria 2021). In practice, this results in a set of neighborhoods having high amenities as well as high shares of college graduates. The draw of such neighborhoods is likely explained by a combination of opportunities for interaction—a topic to which we return below—as well as the endogenous amenities, either private or public, that are abundant and well-funded in high socioeconomic status neighborhoods.

Other endogenous amenities that respond to and, in turn, attract high-income college graduates are reductions in disamenities of crime and pollution. Ellen, Horn, and Reed (2019), for example, argue that the decline in central city crime from 1990 to 2012 has attracted college graduates, generating further neighborhood change. The feedback effects between amenity growth and spatial sorting patterns are quantitatively large.

Different trends driving gentrification—like rising value of time for those with higher incomes and demand for knowledge workers—have twice the impact on spatial sorting when accounting for the indirect effect of endogenous amenities on top of the direct effect of the initial catalyst (Su 2022; Berkes and Gaetani 2019).

Increasing Living Costs and Gentrification

The purchasing power of richer households not only transforms neighborhood amenities but, given housing supply constraints, also drives up the cost of housing and commercial real estate. As higher-income households moved to supply-constrained city centers, some downtown neighborhoods have become less affordable for low-income households and perhaps also less attractive to these households as businesses catering to a lower-income clientele exit, and businesses catering to the tastes of the new, richer and more educated residents enter.

This "gentrification" process has attracted a lot of attention in the popular press for its potentially negative impacts on incumbent residents. Some of the new amenities in gentrifying neighborhoods bring positive value for all residents, including low-income incumbent households; for example, improvements in school quality, reductions in crime, and the entry of businesses that cater to a wide clientele (like grocery stores). Children exposed to gentrification in the suburbs see improvements in credit outcomes later in life, an effect which Baum-Snow, Hartley, and Lee (2019) posit operates mostly through public schools.

However, incumbent low-income residents tend to leave gentrifying urban neighborhoods to escape increasing housing costs, and those who leave of course do not benefit from the new amenities (Ding, Hwang, and Divringi 2016; Brummet and Reed 2019; Dragan, Ellen, and Glied 2020). As discussed earlier, some intuition for this shift is shown in panel C of Figure 4, where low-income residents previously living downtown are outbid for housing by new high-income residents and move further from the city center.

It is worth noting that, in practice, gentrification occurs more via adjustments in replacement patterns than via displacement (Ellen and O'Regan 2011). Low-income residents of gentrifying neighborhoods are, in general, no more likely to move than low-income residents of nongentrifying neighborhoods. What differentiates gentrifying neighborhoods is that when low-income residents do move, they are more likely to be replaced by new high-income residents than by new low-income residents.

The welfare impacts of gentrification are not limited to local residents of gentrifying neighborhoods. The arrival of college graduates' spending power into supply-constrained central city neighborhoods results in increasing housing costs throughout the central city, including in nongentrifying neighborhoods. These spillover effects are large: Couture et al. (2019) and Su (2022) estimate that real income inequality increased by between 20 and 35 percent more than nominal income inequality as a result of gentrification-related growth in housing costs between 1990 and 2010.

Some of this price growth is enjoyed by incumbent households that own properties in gentrifying neighborhoods (Brummet and Reed 2019). However, lower-income households are less likely to own housing. For example, households earning over $100,000 were twice as likely to own a home in 2001 as households earning below $10,000 (Herbert et al. 2005). Moreover, low-income owners may also struggle to afford property tax increases resulting from rising property values (Ding and Hwang 2020; Fu 2022).

Various policies have attempted to maintain housing affordability in gentrifying cities with varying degrees of success. Rent control benefits the households that live in targeted units, but disincentivizes the development of new housing city-wide and, therefore, contributes to higher housing costs in the long-run (Diamond, McQuade, and Qian 2019). Tax incentives for affordable housing development, such as the Low-Income Housing Tax Credit, tend to go towards properties that would have been developed anyway and so barely move the needle on city-wide housing supply and, therefore, on housing costs (Eriksen and Rosenthal 2010). Housing vouchers can support incumbent residents of gentrifying neighborhoods, but much of their value goes to landlords (Collinson and Ganong 2018). New housing development, which can be encouraged by "up-zoning" to allow for more multifamily construction, has been shown to lower housing costs locally (Asquith, Mast, and Reed 2020; Pennington 2021). Community land trusts could potentially allow incumbent households to reap some of the gains from redevelopment and housing price appreciation while maintaining affordability.

Social Exposure and Segregation

Shifts in spatial sorting imply a change in the mix of people residing in different neighborhoods and in opportunities for social interactions. While the immediate benefits of these interactions are hard to measure, a recent literature discussed in this journal by Chyn and Katz (2021) seeks to establish how the demographics of the neighborhood in which a child grows up affects long-run outcomes.[7]

Children who spend more time residing in higher-income commuting zones and counties earn higher incomes, are more likely to attend college, and are less likely to be teen parents later in life (Chetty and Hendren 2018a, b). Although the social mobility benefits of higher-income neighborhoods are partially driven by their amenities, including higher quality schools, lower crime, and lower pollution, social interactions and social capital also play a role. For poor Americans, the number of friends of high socioeconomic status is a robust predictor of social mobility, which is determined in large part by exposure to high income people within residential neighborhoods (Chetty et al. 2022).

[7] Ferreira and Wong (2020) use a survey to help parse the willingness-to-pay for social interactions separately from other physical amenities, such as restaurants.

If social exposure to high economic status groups affects long-run outcomes, then spatial sorting on socioeconomic status may affect intergenerational mobility. This mechanism can exacerbate the growth in nominal income inequality: for example, the sorting response to a permanent shock of a rising labor market premium to skill can amplify the resulting growth in income inequality—the gap between the incomes earned at the top and the bottom of the income distribution—by an additional 30 percent (Fogli and Guerrieri 2019).

These long-term neighborhood effects depend on the demographic composition of a child's immediate neighborhood. Thus, a natural question is whether the urbanization of college graduates is associated with changes in how mixed different neighborhoods are.[8] The sorting patterns described earlier in this paper suggest an ambiguous relationship between changes in neighborhood segregation and changes in the location choice of college graduates. Panel A of Figure 1 and panel B of Figure 2 show a sharp increase in the concentration of white college-educated individuals in certain regions of Core-Based Statistical Areas, especially in areas close to city centers that initially had large shares of non-college-educated and minority inhabitants. Figure 3, however, shows that college-educated newcomers tended to cluster into the same select downtown neighborhoods, instead of evenly mixing with incumbent residents, casting doubt over whether urban revival resulted in more diverse neighborhoods.

To investigate the link, if any, between the urbanization of college graduates and socioeconomic segregation, we return to data for the 100 largest Core-Based Statistical Areas in the United States. We ask whether cities where college graduates urbanized more between 1980 and 2017 also saw decreases in how segregated college graduates were from noncollege graduates across downtown neighborhoods over the same period. We measure the urbanization of college graduates as the share of the college-educated population living downtown, divided by the share of the total population living downtown. There are many ways to measure the degree of local sorting by education. We use two indexes that capture the proximity of noncollege downtown residents to college educated residents within Census tracts. One index measures isolation and the other measures segregation.

We first consider an index of how isolated non-college-educated downtown residents are from college-educated downtown residents. This index is simply equal to the average share of non-college-educated residents in the neighborhoods where non-college-educated individuals live. For example, if a city's isolation index is 0.4, then 40 out of every 100 of an average non-college-educated resident's neighbors are also non-college-educated. Note that a higher index value indicates more isolation

[8] We are not aware of existing research systematically relating urban revival with segregation. In a closely-related inquiry, Freeman (2009) finds that some but not all measures of gentrification are associated with greater diversity within gentrifying neighborhoods.

and that one minus the isolation index captures noncollege residents' exposure to college graduates.[9]

In panel A of Figure 5, the horizontal axis shows the change in the relative urbanization of college graduates from 1980 to 2017. Specifically, this relative urbanization is the share of the college-educated population living downtown divided by the share of the general population living downtown. The vertical axis shows the isolation of downtown non-college-educated residents from college-educated residents, based only on where they live. Isolation of the noncollege educated living downtown declined in all but one of the largest 100 Core-Based Statistical Areas, as shown by the negative numbers on the vertical axis. Looking across metro areas, we find a strong negative relationship between changes in the urbanization of college graduates on the x-axis and changes in the isolation of non-college-educated downtown residents on the y-axis. In other words, between 1980 and 2017, non-college-educated downtown residents experienced rising exposure to college graduates within their residential neighborhoods, and more so in cities with higher rates of college urbanization.

These patterns are perhaps not surprising. After all, aggregate increases in college attainment rates left non-college-educated individuals with fewer peers to live amongst in general (thus the nationwide decreases in the noncollege isolation index). In addition, large influxes of college graduates downtown also left non-college-educated downtown residents with fewer peers to live amongst (thus the negative correlation between the city-level isolation index and college urbanization rates).

To control for these mechanical relationships, we turn to our second measure of local sorting, the η^2 segregation index. The η^2 index normalizes the isolation index to account for changes in the overall share of non-college-educated downtown residents.[10] A higher η^2 index value indicates more segregation. Panel B of Figure 5 compares the change the downtown η^2 segregation index across cities with different degrees of college urbanization. We find a positive and significant relationship between changes in college-educated urbanization and the downtown η^2 index. This finding indicates that the negative relationship between college urbanization and noncollege isolation in panel A of Figure 5 is mechanically driven by aggregate increases in educational attainment rather than local sorting patterns. As seen in the maps for Philadelphia and Chicago in Figure 3, the urbanization of college

[9] The isolation index is the average share of noncollege residents in a Census tract, weighted by the fraction of the city's noncollege residents in each tract. The formula to calculate the index is

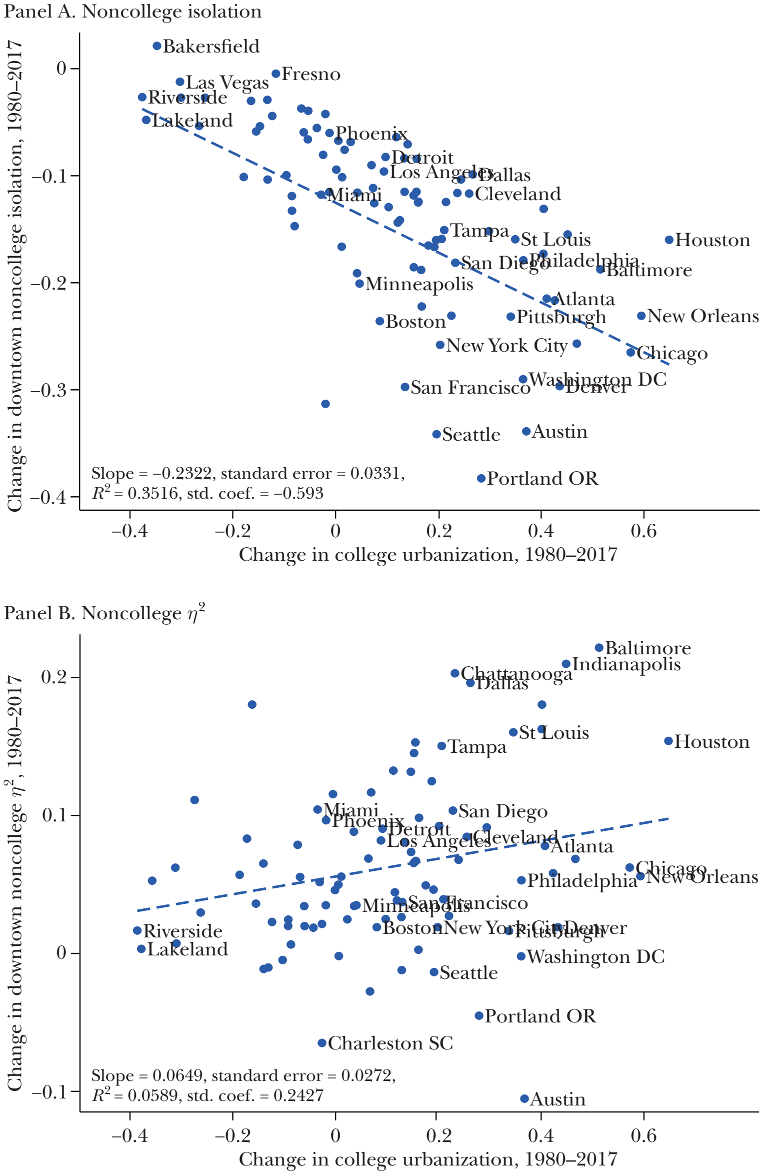
$$\sum_j \frac{C_j}{C_{total}} \cdot \frac{C_j}{C_j + NC_j}$$

where C_j is the number of college graduates in Census tract j, NC_j is the number of non-college residents in tract j, and C_{total} is the number of college graduates in the city.

[10] Specifically, the η^2 index, using the notation in Graham (2018), is equal to $\frac{I - Q}{1 - Q}$, where I is the isolation index, Q is the overall share of non-college-educated in the population (that is, the isolation index under perfect integration) and 1 is the value of the isolation index under perfect segregation. So $I - Q$ is the excess isolation over perfect segregation, normalized by the excess isolation over perfect integration.

Figure 5

Changes in College Urbanization and Downtown Segregation, 1980–2017



Panel A. Noncollege isolation

Slope = –0.2322, standard error = 0.0331, R^2 = 0.3516, std. coef. = –0.593

Panel B. Noncollege η^2

Slope = 0.0649, standard error = 0.0272, R^2 = 0.0589, std. coef. = 0.2427

Source: NHGIS Census (1980, 1990, 2000) & American Community Survey (2008–2012, 2015–2019) (Manson et al. 2022); Longitudinal Tract Data Base (Logan, Xu, and Stults 2014); Holian (2019).

Note: This figure plots changes in college urbanization and changes in segregation in the downtowns of the largest 100 cities, as ranked by Core-Based Statistical Area total population in 2000. College urbanization is the share of the CBSA's college-educated population that lives downtown divided by the share of the CBSA's total population that lives downtown. Downtown is defined as the tracts closest to the center city that make up 10 percent of the CBSA population. The dashed line shows the results of a linear regression of change in the downtown noncollege segregation index, either isolation or η^2, on the change in college urbanization, weighted by city population. The coefficient, standard error, R^2, and standardized coefficient of the regression are reported beneath each panel. The largest 25 cities and outliers are labelled.

graduates into downtowns was distributed in such a way as to generate more sorting by education, not less.[11]

To summarize, as the college-educated have moved downtown, the average non-college-educated resident experiences higher within-neighborhood exposure to college graduates. The normalized η^2 measure of segregation, however, suggests that if the share of college graduates had not been rising for the population as a whole, there could instead have been slightly rising segregation of non-college-educated downtown residents. The η^2 measure of segregation has no micro-foundation in terms of individual preferences, but these results are at least inconsistent with the notion that the young, white, college-educated individuals who drive urban revival were preferentially mixing with non-college-educated residents.[12]

Conclusion

The urban revival of the last few decades reversed a long-term pattern of suburbanization for the college-educated. It represents a sharpening in sorting by education within cities, driven almost entirely by younger cohorts of white college graduates. As college graduates moved downtown, neighborhood amenities evolved to match their tastes, and rising house prices hurt local incumbent residents.

Given the relationship between neighborhood demographics and social mobility, urban revival has, in theory, the potential benefit of exposing urban residents to more educated neighbors. The exposure of non-college-educated individuals to college graduates within residential neighborhoods did rise in gentrifying downtowns. However, the tendency of college graduates to cluster into select downtown neighborhoods, instead of spreading evenly across neighborhoods, likely limited opportunities for interactions across educational lines.

Does the post-pandemic era herald the beginning of another broad shift in sorting patterns? One force that pulled college graduates into city centers over the past 40 years—the rising time costs of commuting—has vanished for a substantial number of college-educated workers with the rise of remote work. In the basic model we proposed, the ability to work from home reduces one's willingness to pay

[11] We also replicate (not shown) panel B using changes in a dissimilarity index instead of the η^2 measure, and we do not find any significant relationship with changes in the urbanization of college graduates.

[12] In Appendix Figure A.5, we replicate Figure 5 but measuring non-college-educated segregation in the suburbs instead of downtown. In the suburbs, the associations between college urbanization and both measures of segregation are not statistically significant and at least one order of magnitude smaller than the associations between college urbanization and segregation downtown. Appendix Figure A.6 shows the association between college urbanization and Black segregation instead of non-college-educated segregation. We find that the urbanization of college graduates is associated with a decline in the isolation of Black downtown residents, but smaller in magnitude than the decline in non-college-educated isolation. The η^2 index for Black residents does not decline downtown or at the CBSA level. We conclude that urban revival is unlikely to be an important factor driving the decline in Black segregation observed since 1970, which Vigdor and Glaeser (2012) attribute to reforms in government practice such as lending discrimination and changing racial attitudes.

for downtown living, because it reduces the need to commute to city centers for work. In this sense, a rising propensity to remote work can be thought of as another change in transportation technology. If this "remote" transportation technology is disproportionately available to the college-educated (Bartik et al. 2020; Dingel and Neiman 2020; Mongey, Pilossoph, and Weinberg 2021), then it disproportionately pushes them away from city centers and towards the suburbs. Davis, Ghent, and Gregory (2021), Delventhal and Parkhomenko (2020), and Duranton and Handbury (2023) provide more complete treatments of what caused the recent rise in remote work, and of the effect that remote work might have on spatial sorting in the future.

Some of the other forces we have discussed in this article are also shifting. Urban rates of violent crime are rising. The key demographics driving the college urbanization—young and white groups—are declining in size. The growing-older and minority demographic groups likely still have a preference for suburban living. That said, downtowns retain their advantage in reducing travel costs to amenities and to other people. Remote work is unlikely to change the fact that higher-income people have stronger tastes for urban amenities and higher value of time. The strength of assortative matching and the appeal of urban venues as meeting places for dating and friendship amongst the young and educated is unlikely to diminish. At the tail end of the pandemic, it is still too early to tell whether there will be another reversal in sorting patterns over the coming decades.

■ *We thank Ryan Hastings and William Zheng for outstanding research assistance. Victor Couture thanks the Centre for Innovative Data in Economics Research at UBC for financial support. Jessie Handbury thanks the Zell/Lurie Real Estate Center for financial support.*

References

Asquith, Brian J., Evan Mast, and Davin Reed. 2020. "Supply Shock versus Demand Shock: The Local Effects of New Housing in Low-Income Areas." Federal Reserve Bank of Philadelphia Working Paper 20–07.

Bartik, Alexander W., Zoe B. Cullen, Edward L. Glaeser, Michael Luca, and Christopher T. Stanton. 2020. "What Jobs Are Being Done at Home during the COVID-19 Crisis? Evidence from Firm-Level Surveys." NBER Working Paper 27422.

Baum-Snow, Nathaniel. 2007. "Did Highways Cause Suburbanization?" *Quarterly Journal of Economics* 122 (2): 775–805.

Baum-Snow, Nathaniel, and Daniel Hartley. 2020. "Accounting for Central Neighborhood Change, 1980–2010." *Journal of Urban Economics* 117: 103228.

Baum-Snow, Nathaniel, Daniel A. Hartley, and Kwan Ok Lee. 2019. "The Long-Run Effects of Neighborhood Change on Incumbent Families." CESifo Working Paper 7577.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy* 115 (4): 588–638.

Berkes, Enrico, and Ruben Gaetani. 2019. "Income Segregation and Rise of the Knowledge Economy." Rotman School of Management Working Paper 3423136.

Boustan, Leah Platt. 2010. "Was Postwar Suburbanization 'White Flight'? Evidence from the Black Migration." *Quarterly Journal of Economics* 125 (1): 417–43.

Brueckner, Jan K., Jacques-François Thisse, and Yves Zenou. 1999. "Why Is Central Paris Rich and Downtown Detroit Poor? An Amenity-Based Theory." *European Economic Review* 43 (1): 91–107.

Brummet, Quentin, and Davin Reed. 2019. "The Effects of Gentrification on the Well-Being and Opportunity of Original Resident Adults and Children." Federal Reserve Bank of Philadelphia Working Paper 19–30.

Chetty, Raj, and Nathaniel Hendren. 2018a. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects." *Quarterly Journal of Economics* 133 (3): 1107–62.

Chetty, Raj and Nathaniel Hendren. 2018b. "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates." *Quarterly Journal of Economics* 133 (3): 1163–1228.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong et al. 2022. "Social Capital I: Measurement and Associations with Economic Mobility." *Nature* 608 (7921): 108–21.

Chyn, Eric, and Lawrence F. Katz. 2021. "Neighborhoods Matter: Assessing the Evidence for Place Effects." *Journal of Economic Perspectives* 35 (4): 197–222.

Collinson, Robert, and Peter Ganong. 2018. "How Do Changes in Housing Voucher Design Affect Rent and Neighborhood Quality?" *American Economic Journal: Economic Policy* 10 (2): 62–89.

Couture, Victor, and Jessie Handbury. 2020. "Urban Revival in America." *Journal of Urban Economics* 119: 103267.

Couture, Victor, and Jessie Handbury. 2023. "Replication data for: Neighborhood Change and the Urbanization of College Graduates." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E186383V1.

Couture, Victor, Cecile Gaubert, Jessie Handbury, and Erik Hurst. 2019. "Income Growth and the Distributional Effects of Urban Spatial Sorting." NBER Working Paper 26142.

Davis, Morris A., Andra C. Ghent, and Jesse M. Gregory. 2021. "The Work-from-Home Technology Boon and Its Consequences." NBER Working Paper 28461.

Delventhal, Matt, and Andrii Parkhomenko. 2020. "Spatial Implications of Telecommuting." Unpublished.

Diamond, Rebecca and Cecile Gaubert. 2022. "Spatial Sorting and Inequality." *Annual Review of Economics* 14: 795–819.

Diamond, Rebecca, Tim McQuade, and Franklin Qian. 2019. "The Effects of Rent Control Expansion on Tenants, Landlords, and Inequality: Evidence from San Francisco." *American Economic Review* 109 (9): 3365–94.

Ding, Lei, and Jackelyn Hwang. 2020. "Effects of Gentrification on Homeowners: Evidence from a Natural Experiment." *Regional Science and Urban Economics* 83: 103536.

Ding, Lei, Jackelyn Hwang, and Eileen Divringi. 2016. "Gentrification and Residential Mobility in Philadelphia." *Regional Science and Urban Economics* 61: 38–51.

Dingel, Jonathan I., and Brent Neiman. 2020. "How Many Jobs Can Be Done at Home?" *Journal of Public Economics* 189: 104235.

Dragan, Kacie, Ingrid Gould Ellen, and Sherry Glied. 2020. "Does Gentrification Displace Poor Children and Their Families? New Evidence from Medicaid Data in New York City." *Regional Science and Urban Economics* 83: 103481.

Duranton, Gilles, and Jessie Handbury. 2023. "Covid and Cities, Thus Far." Unpublished.

Ellen, Ingrid Gould, and Katherine M. O'Regan. 2011. "How Low Income Neighborhoods Change: Entry, Exit, and Enhancement." *Regional Science and Urban Economics* 41 (2): 89–97.

Ellen, Ingrid Gould, Keren Mertens Horn, and Davin Reed. 2019. "Has Falling Crime Invited Gentrification?" *Journal of Housing Economics* 46: 101636.

Eriksen, Michael D., and Stuart S. Rosenthal. 2010. "Crowd Out Effects of Place-Based Subsidized Rental Housing: New Evidence from the LIHTC Program." *Journal of Public Economics* 94 (11–12): 953–66.

Ferreira, Fernando V., and Maisy Wong. 2020. "Estimating Preferences for Neighborhood Amenities under Imperfect Information." NBER Working Paper 28165.

Fogli, Alessandra, and Veronica Guerrieri. 2019. "The End of the American Dream? Inequality and

Segregation in US Cities." NBER Working Paper 26143.

Freeman, Lance. 2009. "Neighbourhood Diversity, Metropolitan Segregation and Gentrification: What Are the Links in the US?" *Urban Studies* 46 (10): 2079–2101.

Fu, Ellen. 2022. "The Financial Burdens of Property Taxes: Evidence from Philadelphia." Unpublished.

Glaeser, Edward L. 2020. "Infrastructure and Urban Form." NBER Working Paper 28287.

Glaeser, Edward L., Matthew E. Kahn, and Jordan Rappaport. 2008. "Why Do the Poor Live in Cities? The Role of Public Transportation." *Journal of Urban Economics* 63 (1): 1–24.

Graham, Bryan S. 2018. "Identifying and Estimating Neighborhood Effects." *Journal of Economic Literature* 56 (2): 450–500.

Guerrieri, Veronica, Daniel Hartley, and Erik Hurst. 2013. "Endogenous Gentrification and Housing Price Dynamics." *Journal of Public Economics* 100: 45–60.

Gupta, Abhimanyu, and Jonathan Halket. 2021. "Household Sorting in an Ancient Setting." Centre for Microdata Methods and Practice Working Paper 01/21.

Heblich, Stephan, Stephen J. Redding, and Daniel M. Sturm. 2020. "The Making of the Modern Metropolis: Evidence from London." *Quarterly Journal of Economics* 135 (4): 2059– 2133.

Herbert, Christopher E., Donald R. Haurin, Stuart S. Rosenthal, and Mark Duda. 2005. *Homeownership Gaps among Low-Income and Minority Borrowers and Neighborhoods.* Washington, DC: US Department of Housing and Urban Development. https://www.huduser.gov/publications/pdf/homeownershipgapsamonglow-incomeandminority.pdf.

Holian, Matthew J. 2019. "Where Is the City's Center? Five Measures of Central Location." *Cityscape* 21 (2): 213–26.

Jackson, Kenneth T. 1987. *Crabgrass Frontier: The Suburbanization of the United States.* New York: Oxford University Press.

Lee, Sanghoon, and Jeffrey Lin. 2018. "Natural Amenities, Neighbourhood Dynamics, and Persistence in the Spatial Distribution of Income." *Review of Economic Studies* 85 (1): 663–94.

LeRoy, Stephen F., and Jon Sonstelie. 1983. "Paradise Lost and Regained: Transportation Innovation, Income, and Residential Location." *Journal of Urban Economics* 13 (1): 67–89.

Logan, John R., Zengwang Xu, and Brian J. Stults. 2014. "Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database." *Professional Geographer* 66 (3): 412–20.

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. 2022. "IPUMS National Historical Geographic Information System: Version 17.0." IPUMS (accessed September 2, 2022).

Margo, Robert A. 1992. "Explaining the Postwar Suburbanization of Population in the United States: The Role of Income." *Journal of Urban Economics* 31 (3): 301–10.

Mieszkowski, Peter, and Edwin S. Mills. 1993. "The Causes of Metropolitan Suburbanization." *Journal of Economic Perspectives* 7 (3): 135–47.

Mongey, Simon, Laura Pilossoph, and Alexander Weinberg. 2021. "Which Workers Bear the Burden of Social Distancing?" *Journal of Economic Inequality* 19 (3): 509–26.

Moreno-Maldonado, Ana, and Clara Santamaria. 2021. "Delayed Childbearing and Urban Revival." Unpublished.

Moretti, Enrico. 2012. *The New Geography of Jobs.* New York: Houghton Mifflin Harcourt.

Pennington, Kate. 2021. "Does Building New Housing Cause Displacement? The Supply and Demand Effects of Construction in San Francisco." Unpublished.

Redfearn, Christian L. 2007. "The Topography of Metropolitan Employment: Identifying Centers of Employment in a Polycentric Urban Area." *Journal of Urban Economics* 61 (3): 519–41.

Small, Kenneth A., and Erik T. Verhoef. 2007. *The Economics of Urban Transportation.* Abingdon, UK: Routledge.

Su, Yichen. 2022. "The Rising Value of Time and the Origin of Urban Gentrification." *American Economic Journal: Economic Policy* 14 (1): 402–39.

US Department of Transportation Federal Highway Administration. 2009. *2009 National Household Travel Survey.* Washington DC: US Department of Transportation.

Vigdor, Jacob L., and Edward L. Glaeser. 2012. *The End of the Segregated Century: Racial Separation in America's Neighborhoods, 1890–2010.* New York: Manhattan Institute for Policy Research.

Journal of Economic Perspectives—Volume 37, Number 2—Spring 2023—Pages 53–74

Constraints on City and Neighborhood Growth: The Central Role of Housing Supply

Nathaniel Baum-Snow

Two distinct patterns have emerged among groups of metropolitan areas and neighborhoods in the United States since 1980. While some have experienced rapid growth in population and new housing units, others have experienced rapid growth in housing prices instead. Between 1980 and 2018, the number of occupied housing units in US metropolitan areas grew from 80 million to 122 million. The location of this growth in housing stock is disproportionately oriented toward smaller, less-dense cities and neighborhoods rather than high-population metropolitan areas. Low-density suburbs have grown much faster than all other types of neighborhoods, accounting for 78 percent of the growth in housing units in the set of neighborhoods observed in 1980. However, home price growth has been fastest in high-density suburbs and the most prosperous and high-density areas of central cities. Moreover, the rate of new housing construction has been low or falling in all types of locations since 2000, but particularly so in the low-density suburbs and rural areas where most of the recent quantity growth has occurred. Altogether, the US housing stock has been getting older, more crowded, and less affordable in recent years.

The net change in housing units in areas already built-up around 1980 has been close to zero, with a constant and low rate of new construction that is barely enough to replace housing that has depreciated beyond livability. In many of these jurisdictions, the existing land use and zoning regulatory environment makes

■ *Nathaniel Baum-Snow holds the Premier's Research Chair in Productivity and Competitiveness and is a Professor of Economic Analysis and Policy at the Rotman School of Management, University of Toronto, Toronto, Canada. His email address is Nate.Baum.Snow@rotman.utoronto.ca.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.53.

it almost impossible to densify. Land available for development only exists at the fringes of metropolitan areas in jurisdictions with low levels of land use regulation. Own-price elasticities of housing supply—the responsiveness in the quantities of housing supplied to changes in housing prices—have experienced secular declines in virtually every type of heavily populated location over the past 40 years. As a result, growth in housing demand has manifested itself much more as price growth than quantity growth, particularly in the most recent decades. Spatial polarization has arisen between severely supply-constrained and high-cost cities and neighborhoods and the somewhat less supply-constrained areas that have experienced home price growth closer to the overall rate of inflation.

Rapid recent housing price growth in the largest metropolitan areas and the densest neighborhoods has come with declining housing affordability in these areas. These areas are the engines of economic growth, yet high housing costs make them inaccessible to many households. While the current regulatory environment bodes ill for the prospects of improving affordability in these areas, all hope is not lost. Policymakers are coming to recognize the costs of the widespread exclusionary zoning policies that maintain existing population and housing unit densities. Recent policy shifts in several states, most notably in California, seek to encourage densification. As it stands now, however, the low and declining elasticity of housing supply, and the legal prohibitions on densification imposed in many jurisdictions, is of growing concern. A society that affords lower-income households the opportunity to buy into high opportunity locations through the purchase (or rental) of small housing units is quickly becoming something of the past.

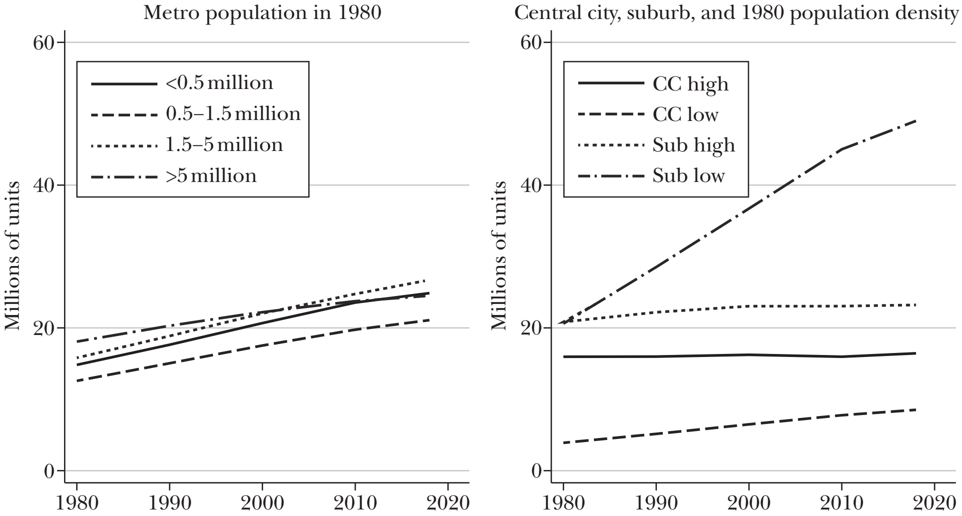
The Evolution of Housing Stocks, Flows, and Prices by Location Type

I begin by describing patterns of housing stocks, construction flows, and prices across the United States since 1980, while not yet considering the mechanisms that may generate these facts. I identify two broad types of locations: those in which housing stocks have increased rapidly but prices have not, and those that exhibit the opposite pattern. The data are clear that new housing construction has become increasingly constrained in high population metro areas and dense neighborhoods, resulting in only slow increases in housing stocks in these locations, along with high rates of price growth. Constraints on population and economic growth in these large metros and dense neighborhoods, which are also the most prosperous, appear to be fundamentally linked to low rates of new housing construction.

Housing Stocks

Figure 1 shows the evolution of the aggregate stock of housing units over time since 1980. The underlying data points are spatial aggregates of decennial US census tract data from 1980, 1990, 2000, and 2010 along with 2008–2012 and 2016–2020 five-year American Community Survey tract data (Geolytics 201u;

Figure 1
Evolution of the Urban Housing Stock in the United States, 1980–2018



Source: Baum-Snow (2023).
Note: The left panel shows the aggregate number of housing units in all metropolitan areas of indicated populations (in millions) for each indicated year. The right panel shows the aggregate number of housing units in all neighborhoods of the indicated types. "CC" denotes central city and "Sub" denotes suburbs. High-density neighborhoods are 2000-definition census tracts of at least 1,000 residents per square kilometer in 1980. All other urban tracts are in low-density neighborhoods.

Manson et al. 2022; US Census Bureau 2011). The left panel has a different line for metropolitan regions of each indicated population size in 1980. The category for greater than five million population has ten partially overlapping regions, including New York, Los Angeles, Chicago, Philadelphia, San Francisco and Washington DC.[1] The next largest category has 21 regions including Atlanta, Cincinnati, Houston, Minneapolis, Pittsburgh, and Tampa. The 0.5-1.5 million population category has 53 regions, including Austin, Buffalo, Dayton, Kansas City, Memphis, and Richmond. The smallest category has 222 regions, including Amarillo, Chattanooga, Des Moines, Fargo, Las Vegas, and Shreveport. The population cutoffs between categories are chosen so that each group of metro regions has roughly the same aggregate stock of housing units. These 306 metropolitan regions are the same as those analyzed in Baum-Snow and Han (2023). They include 76 percent of housing units nationwide in 1980 and 79 percent of housing units nationwide in 2018.

The right panel of Figure 1 has a separate line for each group of census tracts of the indicated type. "CC" denotes central city, which is the unique city containing

[1] The analysis applies appropriate weights for neighborhoods in overlapping metro regions in order to avoid double-counting.

each metropolitan area's central business district. "Sub" denotes suburbs, which includes all census tracts outside of the central city. "High"-density census tracts are those with at least 1,000 residents per square kilometer in 1980 and "low"-density tracts are the remainder. As census tracts are drawn to have (roughly) similar populations, low-density tracts typically have much greater land area than high-density tracts.[2] Each metro region in the data has some high-density central city tracts and low-density suburban tracts. However, larger metros are less likely to have many low-density central city tracts and many smaller metros do not have any high-density suburban tracts.

The left panel of Figure 1 shows that the aggregate housing stock grew faster in the small- and medium-size metropolitan areas than in the largest metropolitan areas. Each of the three smaller types of metropolitan areas experienced an average 1.8 percent growth in housing stock each year during the 1980–2018 period, totaling 68 percent over the full study period. In contrast, the largest metropolitan areas (over 5 million in population) experienced an average annualized 0.9 percent growth in housing stock, totaling just 36 percent over the full period. While these large metropolitan areas collectively hosted more housing units than each of the other three categories in 1980, by 2000 they had been overtaken by two categories of the faster-growing smaller metros (less than 0.5 million and 1.5-5 million in population). The largest metros have been losing population and housing unit share in every decade since 1980 (and perhaps earlier, as well).

While the largest metro areas have been growing at the slowest rates in terms of housing stocks, all types of metro areas have generated fewer new housing units in more recent decades, especially in the 2010–2018 period. This conclusion holds whether growth of housing stock is measured in terms of numbers of raw housing units or as fractions of prior stocks. Each of the smallest three categories experienced less than 1 percent annual growth in stock in the 2010–2018 period, which is a decline from about 2 percent annual growth in the 1980s. This 2010–2018 growth in the stock was even more anemic in the largest category of metro areas at 0.4 percent annual growth, relative to 1.2 percent in the 1980s. Each of the smaller three categories of metro areas had similar stock growth as measured in levels in the 1980s as in the 1990s and declined thereafter. However, the change in the stock in the largest metros declined from an annualized 220,000 units in the 1980s to 200,000 units in the 1990s, 150,000 units in the 2000–2010 period, and 100,000 units in the 2010–2018 period. The slowing pace of new units entering the housing stock is happening in all types of metro areas, but is most pronounced in the largest metro areas.

The right panel of Figure 1 shows that virtually the entire growth in the housing stock occurred in low-density suburban areas, of which the largest metro areas have

[2] The analysis starts in 1980 because this is the first year in which most of the outlying areas near metropolitan fringes were covered by census tracts. However, not all low-density suburban areas included were covered in 1980, resulting in a small undercount of the housing stock for low-density suburbs in 1980 only.

fewer. These areas had a net increase of 28 million housing units from 1980 to 2018; no other type of neighborhood examined managed a net increase of more than five million housing units, though low-density central city neighborhoods did more than double from their low 1980 stock of four million. The fact that growth in low-density suburbs dwarfed growth in other types of neighborhoods in each decade since 1980 is a first hint of a narrative about the recent evolution of the US housing market. As these areas densify and fill up with single family homes, less land becomes available for development, and they enact many of the same land use regulations that have already existed in older more high-density suburbs. This category of neighborhoods cannot keep providing new housing units forever.

Housing Unit Flows and New Construction

Changes in the stock of housing units can come from three sources: new construction adds to the stock; teardowns and abandonment reduce the stock; and renovations of existing buildings affect the stock through their influence on units per building. Focusing on the first source, Figure 2 shows the number of new construction units in each decade prior to the indicated year. For example, counts in 1980 indicate the number of units built 1970–1979, as reported in the 1980 census.[3] Each panel in Figure 2 uses the same samples of census tracts as in Figure 1. Again, the left panel shows a breakdown by metro population in 1980, while the right panel shows a breakdown by central city or suburban high- or low-density neighborhoods in 1980.

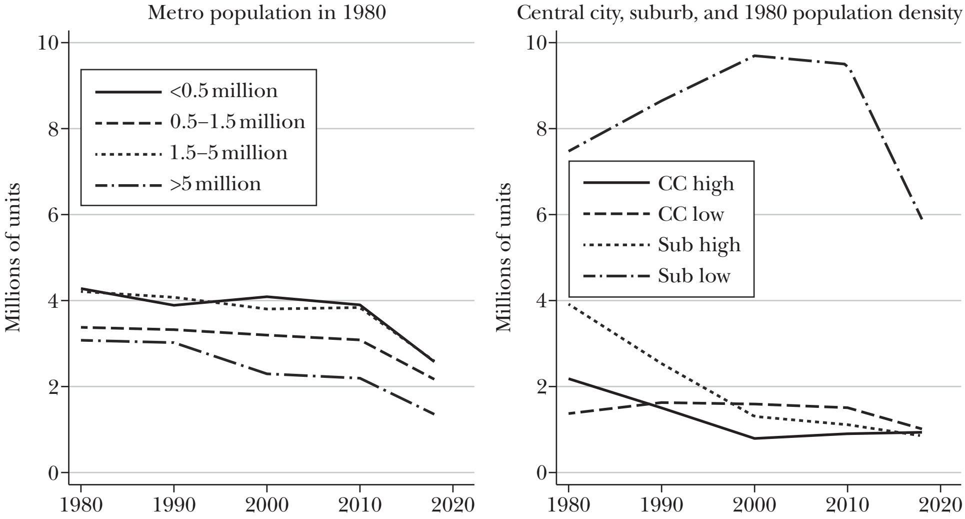
Figure 2 shows that new housing construction has been declining in most locations since 1980. Since 2000, these declines appear in all types of locations but have been particularly strong in the largest metropolitan areas and in low-density suburbs. The left panel of Figure 2 shows that the largest metropolitan areas collectively added about 3 million new units through construction in the 1970s and 1980s: 2.2 million per decade in the 1990–2010 period and 1.4 million on a decadal basis between 2010 and 2018. Comparing to levels in Figure 1, 17 percent of the 1980 housing stock in these metros was from new construction over the prior decade, falling monotonically to just 6 percent of the housing stock in 2018. Among all smaller metropolitan areas, the fraction of the existing stock newly constructed in the prior decade also fell over time, from a much higher 27 percent in 1980 to just 10 percent in 2018. These magnitudes are similar in each of the three smaller metro area size categories.[4]

Commensurate with the fact seen in Figure 1 that most housing stock growth was in low-density suburbs, the right panel of Figure 2 shows that new construction in

[3] Because the decennial census did not ask about the age of housing units in 2010, the counts indicated for 2010 are the 2008–2012 American Community Survey reports of 2000–2009 construction. To make the indicated flows comparable to those for prior decades, the 2018 counts are calculated as (new construction in calendar years 2010–2017)*10/8.

[4] A look at American Community Survey data more disaggregated by year reveals that the rate of housing construction declined precipitously during and immediately after the Great Recession, only recovering to about 75 percent of its 2000–2009 average by 2018.

Figure 2
Housing Unit Construction in Prior Decades, 1980–2018



Source: Baum-Snow (2023).
Note: Plots show the aggregate number of units constructed in the ten years prior to each indicated year in each indicated region. The final data point is for 2018 and reflects eight years of construction inflated to a decadal basis. See the notes to Figure 1 for sample definitions.

this type of neighborhood has dominated that in other types of neighborhoods. For example, in the 2000–2009 period, these neighborhoods added 9.5 million housing units, whereas other types of neighborhoods collectively added only 3.5 million units. Low-density suburban neighborhoods have provided about 70 percent of new housing construction in metropolitan areas in each decade since the 1990s. However, new construction in the booming low-density suburbs has abated quickly since 2010. These neighborhoods are becoming more like their high-density counterparts as they become developed; indeed, each type of neighborhood has experienced a marked decline in the number of new construction units since 2010.

To maintain the growth in housing stocks reported in Figure 1 in the face of declining rates of new construction, it must be that fewer units are being pulled out of the housing stock through teardowns, abandonment, or renovation. Evidence using individual-level property assessment data suggests that renovations have little net effect on the housing stock (Baum-Snow and Han 2023). Thus, it seems likely that this change mostly comes from reductions in teardowns and abandonments. Combining information from Figures 1 and 2, the implication is a teardown/abandonment rate of only about 2 percent per decade in the largest metros since the 1990s, relative to 5 percent in smaller metros. The higher prices in larger cities (documented in the following subsection) may encourage property owners to engage in more robust maintenance, thereby extending the life of properties. Moreover, more recently constructed properties might be of higher quality than those replaced, thereby facilitating declining rates of teardowns and abandonment. When comparing across central city and suburban neighborhoods by density, there is no consistent pattern in teardown/abandonment rates.

While a recent uptick in new construction in dense central city neighborhoods is evident in the data in recent decades (Couture et al. 2019; Couture and Handbury 2020; Baum-Snow and Hartley 2020), this phenomenon is minor relative to prior declines in central city construction rates. High-density central city and suburban neighborhoods both had declining numbers and rates of new construction between 1980 and 2000. High-density suburbs experienced particularly strong declines, with 3.9 million new construction units in the 1970s falling to only 1.3 million new construction units in the 1990s and further declines in subsequent two decades. High-density central city neighborhoods have managed to slightly increase their new construction rates since the low point of a total of about 0.8 million units built in the 1990s. However, this slight growth has been only just enough to replace abandonments and teardowns.

Housing Prices

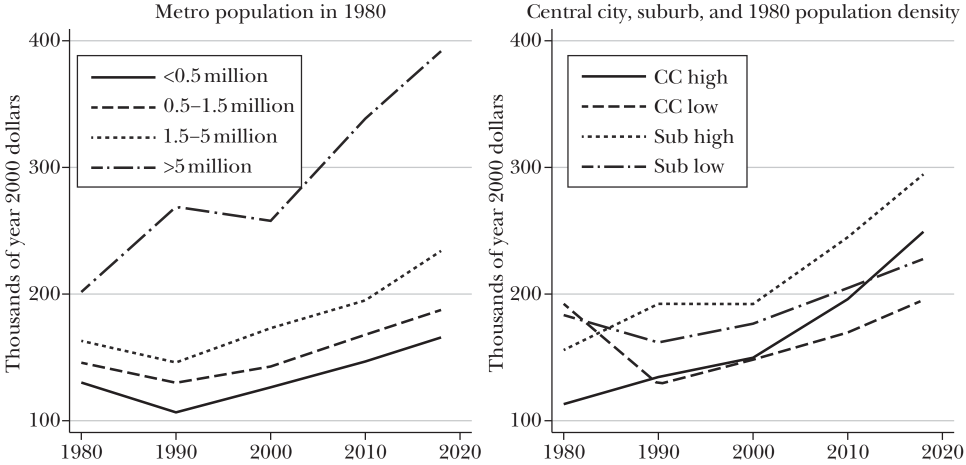
A look at housing prices completes the descriptive picture of how the housing markets in different types of metropolitan areas and neighborhoods have evolved in recent decades. Figure 3 depicts the average self-reported value of owner-occupied housing units from the census or American Community Survey in each indicated year and region, deflated using the Consumer Price Index to year 2000 dollars.[5]

The left panel of Figure 3 shows clear gaps in both levels and growth rates of home prices between metropolitan areas of different sizes. There is a strong positive relationship between the price of housing units and metro population, helping to justify the much lower rates at which housing units are withdrawn from the stock in the larger metros. Moreover, housing prices have risen more rapidly in the largest metros, likely reflecting at least in part the higher costs of building new housing in these locations. In particular, in the category for the largest metros, housing prices increased by 94 percent between 1980 and 2018, relative to 44 percent for metros of 1.5–5.0 million people in 1980, and about 28 percent in each of the two smallest categories of metro area.

The right panel of Figure 3 shows markedly different price growth trends by type of neighborhood. Since 2000, high-density neighborhoods have had much more rapid price growth than low-density neighborhoods, after experiencing similar price growth during the 1990s. Over the full study period, prices in central city high-density neighborhoods grew by 120 percent and those in high-density suburbs grew by 89 percent. However, those in low-density central city neighborhoods hardly changed, while those in low-density suburbs grew by only 24 percent.

[5] While some treatments use a price index rather than prices, to account for housing unit quality such an adjustment makes little difference for measuring the overall trends in the data depicted in Figure 3.

Figure 3
Average Self-Reported Home Values



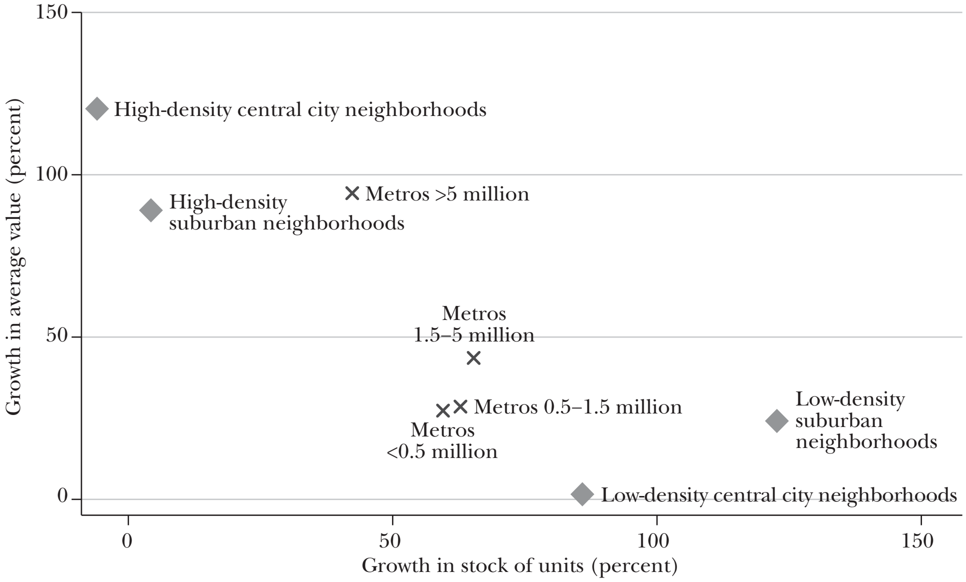
Source: Baum-Snow (2023).
Note: Plots indicate average self-reported home values for owner-occupants. See the notes to Figure 1 for sample definitions.

These changes have been accompanied by lower rates of new construction in the high price growth neighborhoods, as they have less land available for development.

Summary of Empirical Patterns

In summary, two distinct types of locations are clear in the data. Many large metro areas have experienced rapid price growth but anemic quantity growth in the housing market. These large metros have the lowest rates of new construction and the most rapid declines in rates of new construction since 1980. Substantially but not entirely overlapping are the densest suburban and central city neighborhoods that exhibit similar patterns of price and quantity changes. These are sometimes called the "superstar" cities and neighborhoods (Gyourko, Mayer, and Sinai 2013). Examples of superstar cities include San Francisco and Boston. The other type is low-density suburban neighborhoods, where most new construction beyond replacement has occurred and where prices have had relatively low growth rates. Exurbs in the Washington DC, Dallas, Los Angeles, Atlanta, and Houston metro areas fit this profile; each supplied more than one million new housing units between 1980 and 2018. Low-density urban neighborhoods account for a much smaller fraction of the housing stock but have seen some modest net growth, though their low prices may reflect sufficiently weak demand such that prices remain below replacement cost in many such areas.

Figure 4
Growth in Housing Unit Average Values and Stocks, 1980–2018



Source: Baum-Snow (2023).
Note: This figure shows the relationship between the percent growth in average self-reported home values (vertical axis) and the percent growth in the aggregate stock of housing units (horizontal axis) for indicated areas. Calculations use data for 2018 and 1980 reported in Figures 1 and 3.

Figure 4 graphically summarizes these facts. It plots relationships between the 1980–2018 growth rates in prices on the vertical axis against housing units on the horizontal axis for the four categories of metro area and the four types of neighborhood considered above. Metro area categories are shown as x's and neighborhood types are shown as diamonds. There is a cluster of three points in the upper left of the graph, indicating anemic quantity growth and high price growth, that includes high-density neighborhoods and the largest metros. These are the "superstars." The remaining areas, with relatively robust quantity growth but low price growth, appear toward the bottom right of the graph.

Now that some key facts have been established, the following section develops a conceptualization of the forces driving these patterns in the data.

Forces Driving City and Neighborhood Growth

For any good, including housing, understanding observed changes in quantities and prices comes down to isolating forces driving changes in the supply and demand for the good. Once we understand the potential mechanisms driving patterns of changes in housing quantities and prices in the data, we will be in a better position to diagnose why the housing stock has stopped growing in most places, new construction rates have fallen, and prices have risen so quickly in high-density neighborhoods and large cities.

Links from the Labor Market to Housing Demand

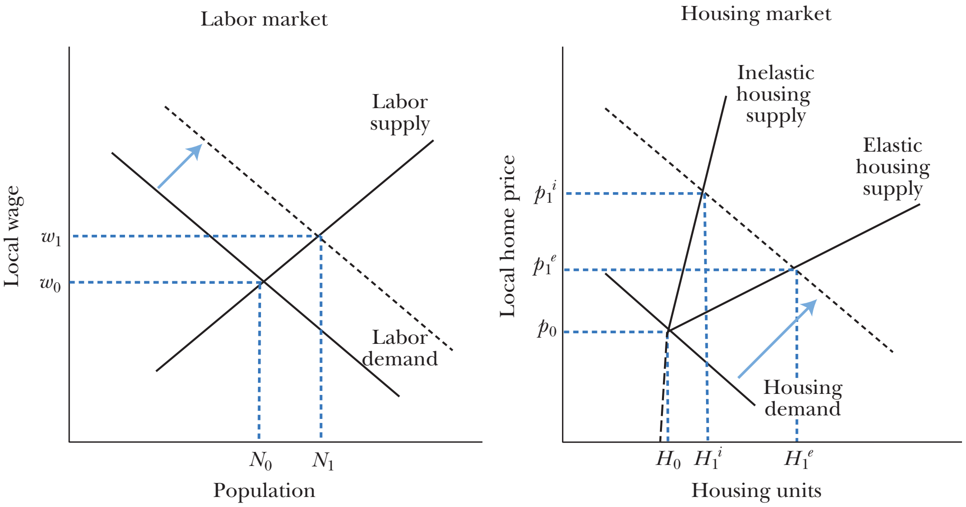
The housing demand curve describes the quantity of housing units that households are willing to buy or rent at each price at a point in time. We can define housing demand for any spatial unit for which data can conveniently be constructed. When each location is thought of as a separate housing market, the drivers of housing demand growth, including population and income growth, can be understood by examining the labor market in these same locations and any other locations within reasonable commute times.

The labor demand curve describes the relationship between the quantity of people firms are willing to hire in a location and the earned income (wage) firms will need to pay in order to hire at that location. As the wage declines, firms are willing to hire more workers. If firms experience an increase in productivity, meaning the amount of output generated by each worker measured either in quantity or dollar terms, firms look to hire more workers at the same wage. This dynamic is the primary reason that certain locations experience outward shifts in labor demand.

While firms in most locations have experienced productivity growth over time, those in certain cities and neighborhoods have experienced particularly strong productivity growth in a way that tends to be correlated with both the industry mix and the skill composition of the workforce. For example, the economies of the Austin, Boston, and San Francisco metropolitan areas have large shares of workers in the high-skilled service sectors, which include professional services, finance, insurance real estate, legal, and accounting services. About 23 percent of employment in these three metro areas was in high-skilled services in 1990, putting each of them in the top 10 percent of metros by this measure. These industries have experienced large productivity gains since 1980 and have employed an increasing share of highly educated and highly paid workers. As a result, labor demand in these metro areas has shifted out by a large amount in recent decades, relative to other locations. Similarly, firms in many high-density suburban areas, including Silicon Valley in suburban San Francisco and Route 128 in suburban Boston, have experienced strong productivity growth, thereby also increasing labor demand particularly rapidly in these areas. At the other end of the spectrum, metro areas more specialized in low-growth industries, most notably manufacturing, have experienced anemic labor demand growth in recent decades. These include the Flint, Gary, and Youngstown metro areas, which each had at least 26 percent of their local employment in manufacturing in 1990 and were above the 85th percentile in the metro area distribution of manufacturing share of employment. Such low productivity growth metro areas are more likely to have lower population levels and densities than their high productivity growth counterparts.

The labor supply curve describes the number of people who are willing to live and work in a location given each level of earnings (or wages). As earnings

Figure 5
Labor and Housing Market Linkages—How Responses to a Productivity Shock Depend on Housing Supply Elasticity



Source: Author's sketch.
Note: These graphs are a stylized depiction of the labor and housing markets in a location.

rise, more people move to a location to take advantage of this higher pay, which is reflected in upward-sloping labor supply. Labor supply conditions differ across cities, in part because of their differences in consumer amenities. All else equal, people are willing to accept a lower wage to live in locations with higher amenities. As such, a rise in the amenity value of a city shifts the labor supply curve out.

Putting labor supply and labor demand together, the left panel in Figure 5 depicts the labor market for a location. Recall that we can define "location" to be a certain metropolitan area, group of metropolitan areas, or group of neighborhoods by type, as is convenient for modeling purposes. In this graph, the initial equilibrium population is N_0 and the equilibrium wage is w_0. If the firms in this location experience productivity growth, the labor demand curve shifts out to the dotted line, bidding up the equilibrium wage to w_1 and drawing in population such that it rises to N_1. The labor market would also be affected if a location experiences a change in the quality of its consumer amenities. If a location becomes a nicer or more convenient place to live, labor supply would shift out. For example, the increased accessibility of suburbs that came with the construction of highways shifted labor supply out in these locations (Baum-Snow 2020).

A positive productivity or amenity shock in a city has a knock-on effect in the housing market to boost housing demand. Figure 5 depicts the situation given a positive labor demand shock. The inflow of workers into this location (with higher wages as well) boosts housing demand, as depicted in the graph on the right. The productivity shock depicted in the graph on the left feeds through to more people willing to rent or purchase housing units at a given price, thereby shifting housing demand out. Other than productivity shocks, changes to mortgage terms for borrowers, like interest rates, or local consumer amenities would also shift housing demand.

Housing Supply

The housing supply curve depicts the cost of providing an additional unit of housing to the market. For positive housing demand shocks, the supply response depends on the cost of constructing additional units of housing and/or preventing existing housing units from depreciating beyond livability. Markets with more elastic housing supply have costs to provide additional housing units that are increasing with quantity growth at lower rates than markets with less elastic supply. For example, a housing market with a supply elasticity of two means that housing quantities are twice as responsive to the same price change relative to a market with the steeper supply curve with the lower elasticity of one. The right half of Figure 5 depicts examples of elastic and inelastic housing supply curves. The same housing demand shock manifests as greater price growth and smaller quantity growth in environments in which housing supply is less elastic. Because national population and productivity growth leads to secular housing demand growth in most locations, differences in housing supply elasticities across locations shape whether this demand growth results more as rising housing prices or rising populations and more new housing construction.

Because of the irreversibility of housing once constructed, the housing supply curve is often modeled as kinked at a long-run equilibrium point. At quantities and prices below this point, labeled (H_0,p_0) in the right graph of Figure 5, housing supply is very inelastic; that is, the supply of housing already constructed does not decline by much over any short-run time interval. In Figure 5, any price below p_0 is below replacement cost. Reductions in the quantity of housing supplied in the face of a negative housing demand shift would only come through the slow process of depreciation. For this reason, negative housing demand shocks tend to come with small declines in quantities and large declines in prices in all types of locations (Glaeser and Gyourko 2005).

The cost of constructing new housing depends centrally on land availability and land use regulation. Areas in which procurement of land for new construction is easy tend to have higher supply elasticities, allowing these areas to grow quickly with demand growth. The cost of acquiring and preparing new land for development depends on topography, the existing built environment, and regulation. Areas with flat topography, low levels of existing development, and minimal regulatory costs have been found to have relatively elastic housing supply (Baum-Snow and Han 2023; Saiz 2010). Building new housing is more costly in hilly cities like Pittsburgh, Asheville, and Wheeling than in flat cities like Dayton, Oklahoma City, and Wichita. Big cities like New York, Los Angeles, and Chicago tend to be more densely built up, making housing supply less elastic in these locations. More tightly regulated cities like Boston, Philadelphia, and Seattle also tend to have less elastic housing supply.

Housing supply elasticity also differs across neighborhoods within metro areas. Built-up neighborhoods near central business districts tend to have the least elastic housing supply in metro areas. This is likely in part because frictions associated with assembling larger plots of land suitable for multiple residences have been found to make redevelopment very costly and difficult in the most densely developed areas (Brooks and Lutz 2016). Moving away from central business districts into suburban areas and metropolitan fringes, more land becomes available for development as built-up densities fall, thereby allowing housing supply elasticities to rise.

Differences in land use regulation across neighborhoods and jurisdictions also influence variation in local supply elasticities within metro areas. Using municipality level data from the Wharton Residential Land Use Regulatory Index, Baum-Snow and Han (2023) demonstrate that regulation is most restrictive in high-density inner suburbs, compared with other types of municipalities. The Wharton Index is broad-based and considers the housing development delays and additional costs that can be imposed by political actors, court cases, planning reviews, zoning variance requirements, density restrictions, permitting, impact fees, and approval delays. Regulation of housing construction is very complicated and differs in many dimensions across jurisdictions, making the Wharton Index an invaluable summary measure of such regulation.

One very common component of jurisdiction- or neighborhood-level land use regulatory regimes is either a minimum lot size area per housing unit or a maximum floorspace-to-lot-size area ratio restriction. Residents of many neighborhoods support such restrictions, as they help to maintain the local character, in part by excluding lower-income arrivals who could only afford to rent or buy properties that are smaller than allowed. For this reason, such minimum lot size restrictions are often referred to as "exclusionary zoning." Such restrictions tend to impose lower housing densities at the edges of central cities and in inner-ring suburbs than are justified by local housing demand conditions. Moving out from central business districts, increasingly strict regulation competes with the increasing land available for development to influence housing supply elasticities. These observations are consistent with evidence in Figures 2 and 3 that high-density central city and suburban neighborhoods have experienced very low growth in their housing stocks, little new construction, and rapidly rising prices since 2000. Gyourko, Hartley, and Krimmel (2021) document that between 2006 and 2018 no metropolitan area experienced substantially reduced land use regulation, but many jurisdictions have imposed more stringent minimum lot size restrictions in particular.

The earlier empirical observations (especially Figures 1 and 3) have shown relationships between growth in housing quantities and prices. In the largest cities and densest neighborhoods, price growth has been more pronounced than quantity growth. In contrast, other types of locations have featured more rapid quantity growth and less rapid price growth. The most natural explanation for these patterns is that variation in housing supply elasticity is driving the divergence. Moreover, the evidence is consistent with the idea that housing supply elasticities have been declining over time in large cities and dense neighborhoods (as argued in Orlando and Redfearn 2022). Denser residential neighborhoods, especially in the suburbs, commonly implement zoning and land use regulations that lock in the existing level of density. Moreover, redevelopment of the housing stock becomes more difficult as hold-up problems in land assembly bind more with denser development. As the amount of land available for development falls, more stringent zoning restrictions are imposed, and land assembly frictions become more binding, housing supply elasticities will decline. The result is that increasing rates of price growth (seen in the right panel of Figure 3) have gone along with more rapid declines in new construction (Figure 2) and lack of growth in stocks (Figure 1) since 2000.

Margins of Response: Units versus Floorspace

To this point, housing markets have been described in terms of housing units. When considering implications for affordability, housing units is a sensible measure. However, a more general characterization of the market can be framed in terms of units of housing services, which can be proxied by floorspace. This characterization, which additionally incorporates the intensive margin of housing unit quality, is a primary consideration in the housing production literature.

Evidence from single-family homes is that the fraction of developers' costs that go to land is about one-third (Combes, Duranton, and Gobillon 2021). Because other inputs to production have more similar prices across locations, areas in which land is less expensive tend to sprawl more—for example, with more single-level homes. In areas with higher land prices (and stronger demand conditions), properties become more vertical. In addition, properties also start having multiple units given sufficiently high land prices, if this is allowed by zoning regulations. With exclusionary/minimum lot size zoning policies in place, the associated strong housing demand conditions often lead to large single-family homes instead, often in prime locations. Indeed, calculations using the ZTRAX Historical Assessment Database (Zillow 2017) reveal that floorspace in the average new construction unit of housing, which tends to be purchased by higher-income households, rose from 2,235 square feet in 2000 to 2,808 square feet in 2014. Exclusionary zoning policies thus have important negative consequences for affordability and constrain the growth of neighborhoods. If zoning regulations in these areas were relaxed, more housing units could be built per unit of land, allowing for densification. Instead, the market responds to this sort of regulation by providing units that are larger and generally of higher quality than would exist absent regulation.

Evidence on Local Supply Elasticities

Having established that the local housing supply elasticity is a central determinant of the nature of local growth in response to outward shifts in housing demand, this section explores empirical evidence on the variation in the supply elasticities for housing units and floorspace across cities and neighborhoods of different types. To understand the prospects for population growth and new housing development across different types of locations, it then explores how the components of these elasticities that are accounted for by new construction vary by type of location.

Estimation and Identification

The housing supply elasticity is the ratio of the percent change in the quantity of housing supplied by the market in response to a given percent change in the housing price. The supply elasticity is determined only by the actions of housing developers and owners in response to a price change, not by actions of housing consumers. To estimate supply elasticities from the data, one thus needs information on prices and quantities over time, plus a source of variation in changes in housing demand that is unrelated to factors that influence construction costs. By looking at multiple neighborhoods that are observationally equivalent before exogenous housing demand shocks, one can estimate the housing supply elasticity for this type of neighborhood. As in the right panel of Figure 5, by looking across neighborhoods with different attributes that influence supply elasticity, one can estimate the variation in supply elasticity as a function of these supply influencers.

Following the logic of Figure 5, Baum-Snow and Han (2023) use shocks to labor demand in commuting destinations of neighborhood housing markets as the central source of identifying variation for the estimation of neighborhood housing supply elasticities. These labor demand shocks are isolated to be driven by the historical industry composition in commuting-accessible locations, interacted with industry-specific growth rates. This empirical strategy follows in the spirit of Bartik (1991), who recognized that historical industry shares interacted with industry-specific growth rates can be treated, in certain situations, as exogenous productivity shocks. In this empirical setting, census tracts with different sized labor demand shocks in commuting destinations are compared. This variation in labor demand shocks feeds through to variation in housing demand shocks, allowing for recovery of estimates of housing supply.

As a practical matter, estimating housing supply elasticities involves a two-stage least squares or instrumental variable process. The objective is to recover coefficients on price growth rates in regressions of the growth rate in housing quantities on the growth rate in housing prices estimated using census tract level data.[6] The price growth that identifies supply elasticities must be driven by demand shocks only. Therefore, the first stage of the instrumental variables process must predict the neighborhood level growth in housing prices using variation in housing demand only, holding housing construction and maintenance costs constant. The shock to housing demand used in Baum-Snow and Han (2023) comes from labor demand shocks in commuting destinations, as proxied by the earlier industry shares in these

[6] Baum-Snow and Han (2023) use the ZTRAX transactions and assessment data from Zillow (2017), which includes property level information on units and floorspace that is aggregated up to the census tract level in 2000 and 2010. ZTRAX transactions information along with housing attributes are used to construct tract level home price indexes that control for differences in housing unit attributes across tracts and over time.

Table 1
Distributions of Tract Supply Elasticities by Location Type

	Units					*Floorspace*					
	p25	*p50*	*p75*	*mean*	*SD*	*p25*	*p50*	*p75*	*mean*	*SD*	*Tracts*
Panel A. All tracts pooled											
	0.14	0.29	0.44	0.29	0.20	0.33	0.53	0.70	0.51	0.26	50,410
Panel B. By metro region population size											
<0.5 m	0.28	0.44	0.57	0.41	0.20	0.48	0.66	0.78	0.63	0.23	12,489
0.5–1.5 m	0.23	0.36	0.49	0.35	0.19	0.45	0.60	0.74	0.59	0.23	10,630
1.5–5 m	0.14	0.26	0.38	0.25	0.17	0.34	0.49	0.66	0.49	0.24	13,136
>5 m	0.04	0.15	0.29	0.16	0.17	0.17	0.34	0.56	0.36	0.26	14,155
Panel C. By central city or suburb status and 1980 tract population density											
CC High	0.06	0.17	0.28	0.17	0.16	0.22	0.37	0.51	0.36	0.21	11,172
CC Low	0.26	0.40	0.52	0.38	0.20	0.47	0.16	0.73	0.59	0.23	4,245
Sub High	0.06	0.16	0.26	0.16	0.15	0.22	0.36	0.50	0.35	0.22	14,115
Sub Low	0.33	0.43	0.53	0.43	0.16	0.58	0.70	0.80	0.68	0.19	20,878

Source: Supply elasticities are calculated as described in Baum-Snow and Han (2023).
Note: Columns indicate the 25th, 50th, and 75th percentiles, means, and standard deviations of units or floorspace supply elasticities across census tracts in areas indicated at left. Sample definitions are the same as in Figures 1, 2, and 3.

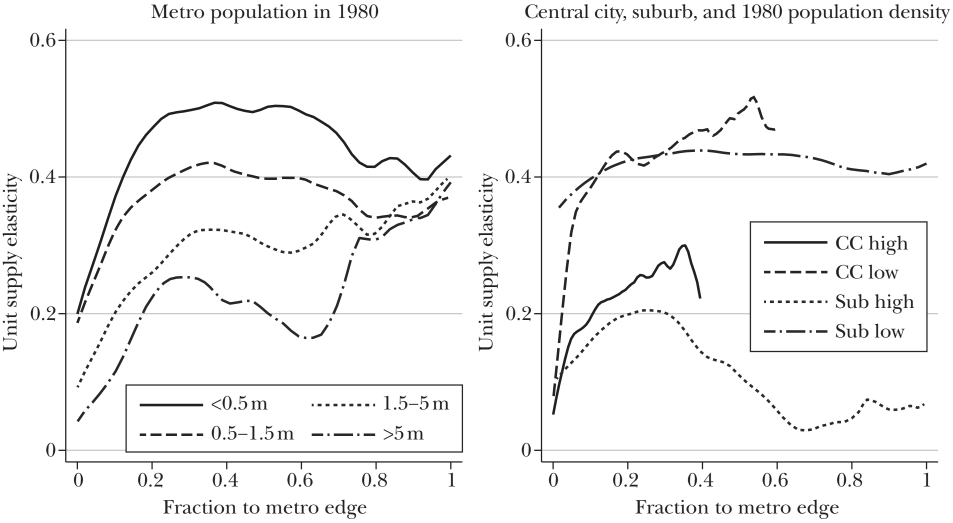
areas interacted with growth rates of industry shares in the economy as a whole. Coefficients on labor demand shocks in this first-stage regression capture exogenously generated changes in housing prices.

The second-stage regression uses these exogenously generated shifts in the price of housing as an explanatory variable in a regression with the change in the quantity of housing as the dependent variable. To recover information on how housing supply elasticities differ by Census tract attributes, home price growth is also interacted with the fraction of the tract land area that is developed in 2001, the distance to the central business district, and the fraction of tract land that is flat. To allow for variation across metropolitan areas of different development intensity, topography, and regulatory environments (as measured by the Wharton Index), elasticities are also allowed to vary by these metropolitan area factors.

Unit and Floorspace Supply Elasticities

Table 1 shows distributions of unit and floorspace supply elasticities estimated using this approach across all urban census tracts from Baum-Snow and Han (2023), along with breakdowns by metropolitan area size and neighborhood type. As seen in panel A, the median metropolitan census tract had a housing unit supply elasticity of 0.29 and a floorspace supply elasticity of 0.53 estimated for the 2000–2010 period. However, there is a fair amount of dispersion in these tract-specific estimates, with standard deviations of 0.20 and 0.26, respectively. The larger floorspace elasticities indicate that as prices rise, housing units tend to become larger, which is consistent with the idea that they will tend to be inhabited by higher-income residents.

Figure 6
Housing Unit Supply Elasticities by Location Type



Source: Baum-Snow and Han (2023).
Note: Plots show mean housing unit supply elasticities estimated in Baum-Snow and Han (2023) across all census tracts in each indicated type of location by an index of distance to metro areas' central business districts. This index is zero at central business districts and one for the furthest tract from it in each metro area.

Supply elasticity estimates in panel B are summarized for groups of metro regions in each indicated population size category. Higher-population metros tend to have smaller supply elasticities. The median tract in metros of at least five million people has a supply elasticity of just 0.15. This number monotonically increases as metro population decreases to 0.44 for the median tract in metropolitan areas of less than 500,000 residents. The greater land available for development in smaller metros along with lower regulation means they have much larger housing supply elasticities. Estimates in panel C show how strongly supply elasticities depend on neighborhood density and location. Low-density city and suburban neighborhoods alike have much higher supply elasticities than do their high-density counterparts. Looking both across cities of different sizes and neighborhoods of different densities, we see that the density of existing development is a key determinant of the ease of expanding the housing stock.

Figure 6 shows how estimated unit supply elasticities differ by distance to the central business district and by metro or neighborhood type. To make metropolitan areas of different shapes and sizes comparable, distance to the central business district is indexed to be between zero and one, where the census tract that is furthest from the central business district gets assigned an index value of one. Because of this indexing scheme and the fact that census tracts are typically much larger near urban fringes, the density of the data declines quickly between 0.6 and one, thereby reducing the precision of supply elasticity estimates in these areas. Commensurate with the evidence in Table 1, denser neighborhoods and more populous metros have lower supply elasticities at all distances from the central business district. Each line in Figure 6 is a local polynomial smoothed regression line of the elasticity estimates reported in Table 1 as a function of distance from the central business district for the indicated set of neighborhoods.

There is a clear pattern of supply elasticities increasing with distance from the central business district in all samples of census tracts shown. Moving further away from the central business district, two countervailing forces affect supply elasticities. First, the amount of land available for development on average increases. Areas near a central business district are usually already heavily developed, making it particularly costly to build new housing. A main reason that average supply elasticities in the higher-population metros are below those in lower-population metros at central business districts, as seen in the left panel of Figure 6, is that such areas are more completely developed in the larger metros. Pushing in the opposite direction, land-use regulations are typically more stringent in high-density suburbs, and especially those dominated by single family homes. Minimum lot size zoning policies are very common in these types of jurisdictions, thereby making densification of housing units difficult and costly.

Plots in the right panel of Figure 6 show that neighborhood population density is highly correlated with supply elasticity. Low-density suburban neighborhoods have remarkably flat supply elasticities in distance from the central business district, at about 0.4 on average. Less densely populated central city tracts have similarly high, or even slightly greater, average supply elasticities, except right at central business districts where few such tracts exist. A striking difference appears, however, between supply elasticities in high-density central city and suburban neighborhoods conditional on distance to the central business district. Beyond about 5 percent of the way to the metro edge, central city neighborhood supply elasticities (solid line) exceed those for suburban neighborhoods (short-dashed line), though both are always below 0.3. This gap may reflect the higher levels of land use regulation in the suburban jurisdictions. In the 137 metro areas for which the central city's 2006 Wharton Index was calculated, the average index for suburban jurisdictions is 0.1 greater than that for the central city, where the index is expressed in standard deviation units.

Components of Unit Supply Elasticities

As home prices rise, homeowners, landlords, and housing developers have incentives to invest in new housing supply in different ways. Housing developers have more opportunities to build new housing profitably, and the associated new housing construction accounts for about 60 percent of the overall unit supply response to price growth (Baum-Snow and Han 2023). In addition, homeowners and landlords of rental properties may respond to price increases by improving quality and/or by subdividing to create additional units. A basement or attic could be converted into a separate apartment. An addition could accommodate a new apartment and/or additional living space. Improved maintenance may allow an old house to remain livable for longer. As home prices rise, teardowns and property abandonment decrease. Supply elasticities incorporate many margins of response beyond new construction. The largest margin of supply response to price signals depends on how costly it is to develop new housing. In areas with little developable land and/or high regulation that limits density, supply responses will be more heavily weighted toward maintenance and reconfiguration of existing structures rather than new construction. Indeed, evidence in Baum-Snow and Han (2023) indicates that redevelopment, while common, is very inelastic to price.

In addition, evidence in Baum-Snow and Han (2023) indicates that new construction tends to represent a greater fraction of unit supply responses to price growth in less dense and less regulated locations. In the largest metros by population, new construction in downtown areas is simply not responsive at all to price signals. In these areas, close to 100 percent of new unit supply is redevelopment of existing properties, but redevelopment is very price insensitive on the margin. In metros of less than 1.5 million people in 1980, about 20 percent of unit supply responses to price growth at central business districts is from new construction. Moving away from central business districts increases the fraction of unit supply elasticities that come from new construction, but it remains lower in more populous and more heavily developed metro areas. At the halfway point from central business districts to metro edges, about 50 percent of supply elasticities in the largest metros are through new construction. This percentage increases monotonically as metro area population falls, such that in the smallest metros about 75 percent of the unit supply elasticity is from new construction at this distance from the central business district.

New construction also contributes to unit supply in different ways across different types of neighborhoods. In central cities, low-density neighborhoods have higher shares at all distances from the central business district, topping out at almost 80 percent in the most peripheral low-density neighborhoods of central cities. High-density central city neighborhoods have lower fractions of supply responses from new construction. But they also rise with distance from the central business districts to reach about 60 percent in these types of peripheral neighborhoods. Interestingly, these fractions for high-density central city neighborhoods exceed those for high-density suburbs at all common distances from the central business district. This pattern may reflect exclusionary zoning policies in such dense suburbs, which are typically stricter than central city land use restrictions; in contrast, the zoning codes in many central cities actively encourage infill and densification. The profile of fraction supply elasticity from new construction with respect to distance from the central business district in lower density suburbs is remarkably similar to that for low-density central city neighborhoods, perhaps reflecting similar regulatory environments.

Discussion and Implications

Under current conditions, the largest metropolitan areas in the United States do not have many neighborhoods with land available for development that are accessible to their urban centers. As a result, in recent decades, more rapid growth has occurred in small- and medium-size metropolitan areas while growth rates in the largest and most productive metropolitan areas have slowed. This pattern of growth has come with serious challenges for facilitating accessibility to affordable housing. Constraints on housing supply exacerbate the spatial polarization generated by variation in growth in labor demand across locations. A set of "superstar cities" offer greater labor market opportunities, but have housing markets that are increasingly expensive to access. Duranton and Puga (2022) document that housing prices at the fringes of the New York, Los Angeles, San Francisco, Washington DC, Boston, Seattle, and San Diego metropolitan areas are more than $200,000 above construction costs, wedges that can only be so large because of land use restrictions. On the other end of the spectrum, cities with slow productivity growth typically have highly accessible housing markets but offer relatively fewer labor market opportunities.

The cross-city dichotomy mentioned above is replicated at the neighborhood level, especially in the largest metropolitan areas. Higher-income neighborhoods, including many inner suburbs, tend to have severe constraints on additional housing supply, which plays a central role in limiting the growth of many of the most productive, high amenity, and high opportunity neighborhoods. The lowest-cost neighborhoods in large metropolitan areas tend to be of two types. Some low-income neighborhoods in central cities and certain suburbs have low opportunity and amenities, with housing prices that are not sufficiently high to justify redevelopment. In such cases, the redevelopment that does occur typically comes with gentrification. The other type is neighborhoods at urban fringes, whose very long typical commutes justify their weaker housing demand.

In theory, one straightforward way to improve affordability and increase housing supply would be to relax exclusionary zoning policies, which are especially prevalent in heavily residential suburban jurisdictions. Even if doing so does not change the price of floorspace, it would allow lower-income households to buy into higher quality neighborhoods through the purchase or rental of smaller housing units. To have much of an effect on the price of floorspace, such reductions in regulation would have to be carried out in a broad-based way across many jurisdictions within a metropolitan area simultaneously. Otherwise, the impacts on aggregate housing supply, and thus floorspace prices, would be small. Considering all metro areas in the United States, Hsieh and Moretti (2019) argue using a quantitative model that national GDP growth would have been 36 percent larger between 1964 and 2009 if the most productive metro areas like New York and San Francisco were to relax their land use restrictions to the national median. The reason is that these most productive regions would be able to host greater populations, thereby increasing output per capita. Within metro areas, the quantitative model in Couture et al. (2019) indicates that the costs of the downtown gentrification faced by incumbent low-income residents due to associated higher housing prices would be mostly mitigated by quadrupling neighborhood housing supply elasticities.

While there is little theoretical controversy that relaxing density restrictions enhances welfare of lower-income households by improving housing affordability, as a practical matter it is rare for such restrictions to be loosened much (Gyourko, Hartley, and Krimmel 2021). Broad-based relaxation of land use restrictions can lead to sufficiently reduced land values (through supply expansion effects) that incumbent owner-occupants are worse off (Duranton and Puga 2022). Moreover, there is a perception that the amenity values of higher-income and more density restricted neighborhoods may deteriorate with densification. Nevertheless, there has been a recent push by some states and large cities to relax density restrictions on neighborhoods. For example, in 2018, the city of Minneapolis passed a law abolishing all single-family zoning, affecting 60 percent of the land area of the city. In 2021, the state of California passed a law (Senate Bill 9) requiring jurisdictions to allow additional low-density housing units to be built in neighborhoods zoned for single families. It is still too early to tell whether these modest relaxations of density limiting zoning requirements have appreciably increased local housing supply elasticities. However, there is evidence of negative amenity effects from a similar policy that was recently implemented in Vancouver (Davidoff, Pavlov, and Somerville 2022).

With relaxation of minimum lot size zoning constraints in established residential neighborhoods probably at best only a partial solution to the affordability challenge, what are other options? The post-pandemic cratering of demand for office real estate (Gupta, Mittal, and Van Nieuwerburgh 2022) provides an opportunity. While direct conversion of buildings from office to residential is mostly impractical, prime land that had been devoted to offices will become available over time with the possibility of dense residential development—provided that zoning rules permit it. Moreover, until recently there was a large amount of land zoned for industrial uses within cities that has been underutilized. These commercial areas are likely to face less local opposition to dense residential construction than many established neighborhoods. Most local zoning ordinances were written at a time when industrial land hosted important shares of local economic activity and have not been updated to accommodate declines in demand for manufacturing and office land uses. As affordability challenges bite more, this is thus a good time for jurisdictions to implement reviews of zoning codes to allow more flexibility in conversions to residential land use.

One source of uncertainty associated with imposing changes in zoning laws is that anticipated welfare consequences are not well understood, except as outputs of quantitative models. As such, additional empirical studies on the impacts of laws that limit the scope of various types of zoning policies would be informative. While theorizing has been successful, we have little credible causal evidence about the extent to which amenity values in neighborhoods that experience densification through relaxed zoning restrictions are affected, and especially why. Research that considers such endogenous amenity responses along with housing market effects in a unified way would be greatly informative. In particular, evaluations of strategies that both enhance affordability and also limit associated negative externalities, especially in conceptual environments that accommodate demand heterogeneity and cross-neighborhood variation in supply conditions, represent an important research agenda.

■ *I thank Erik Hurst, Nina Pavcnik, Heidi Williams, and Timothy Taylor for helpful comments. James Macek provided excellent research assistance.*

References

Bartik, Timothy J. 1991. *Who Benefits from State and Local Economic Development Policies?* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Baum-Snow, Nathaniel. 2020. "Urban Transport Expansions and Changes in the Spatial Structure of U.S. Cities: Implications for Productivity and Welfare." *Review of Economics and Statistics* 102 (5): 929–45.

Baum-Snow, Nathaniel. 2023. "Replication data for: Constraints on City and Neighborhood Growth: The Central Role of Housing Supply." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E185504V1.

Baum-Snow, Nathaniel, and Lu Han. 2023. "The Microgeography of Housing Supply." Unpublished.

Baum-Snow, Nathaniel, and Daniel Hartley. 2020. "Accounting for Central Neighborhood Change: 1980–2010." *Journal of Urban Economics* 117: 103228.

Brooks, Leah, and Byron Lutz. 2016. "From Today's City to Tomorrow's City: An Empirical Investigation of Urban Land Assembly." *American Economic Journal: Economic Policy* 8 (3): 69–105.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon. 2021. "The Production Function for Housing: Evidence from France." *Journal of Political Economy* 129 (10): 2766–816.

Couture, Victor, Cecile Gaubert, Jessie Handbury, and Erik Hurst. 2019. "Income Growth and the Distributional Effects of Urban Spatial Sorting." NBER Working Paper 26142.

Couture, Victor, and Jessie Handbury. 2020. "Urban Revival in America." *Journal of Urban Economics* 119: 103267.

Davidoff, Thomas, Andrey Pavlov, and Tsur Somerville. 2022. "Not in My Neighbour's Back Yard? Laneway Homes and Neighbours' Property Values." *Journal of Urban Economics* 128: 103405.

Duranton, Gilles, and Diego Puga. 2022. "Urban Growth and Its Aggregate Implications." NBER Working Paper 26591.

GeoLytics. 201u. " Neighborhood Change Database, 1970-2010". Somerville, NJ. GeoLytics, Inc.

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

Gupta, Arpit, Vrinda Mittal, and Stijn Van Nieuwerburgh. 2022. "Work from Home and the Office Real Estate Apocalypse." NBER Working Paper 30526.

Gyourko, Joseph, Jonathan S. Hartley, and Jacob Krimmel. 2021. "The Local Residential Land Use Regulatory Environment across U.S. Housing Markets: Evidence from a New Wharton Index." *Journal of Urban Economics* 124: 103337.

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

Hsieh, Chang-Tai, and Enrico Moretti. 2019. "Housing Constraints and Spatial Misallocation." *American Economic Journal: Macroeconomics* 11 (2): 1–39.

Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. 2022. IPUMS National Historical Geographic Information System: Version 17.0 [1980 Census of Population and Housing STF1 Tract 1980_STF1 ds104; 2020 Amerian Community Survey 5 Year Data Tract 2016_2020_ACS5a ds249]. Minneapolis, MN: IPUMS. 2022. http://doi.org/10.18128/D050.V17.0.

Orlando, Anthony W., and Christian L. Redfearn. 2022. "Houston, You Have a Problem: How Large Cities Accommodate More Housing." Unpublished.

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

Zillow. 2017. "ZTRAX: Zillow Transaction and Assessor Dataset, 2017-q4." Zillow Group, Inc. https://www.zillow.com/research/ztrax/.

Quantitative Urban Models: From Theory to Data

Stephen J. Redding

One of the most striking economic features of our world is the uneven distribution of economic activity across space. This concentration is evident in the existence of cities. By 2018, 55 percent of the world's population lived in urban areas, with one in eight urbanites residing in 33 megacities with more than ten million inhabitants (United Nations 2019). But similar concentration is observed within cities as well. Some parts of a city may have access to natural water and be well-suited for heavy industrial use. Other parts of a city may have access to open space and scenic views and be well disposed for residential use. Yet other parts of a city may have good transport connections and be accessible for retail activity. As a pedestrian walks from one city block to another, land use can change sharply from residential to commercial land use and back again.

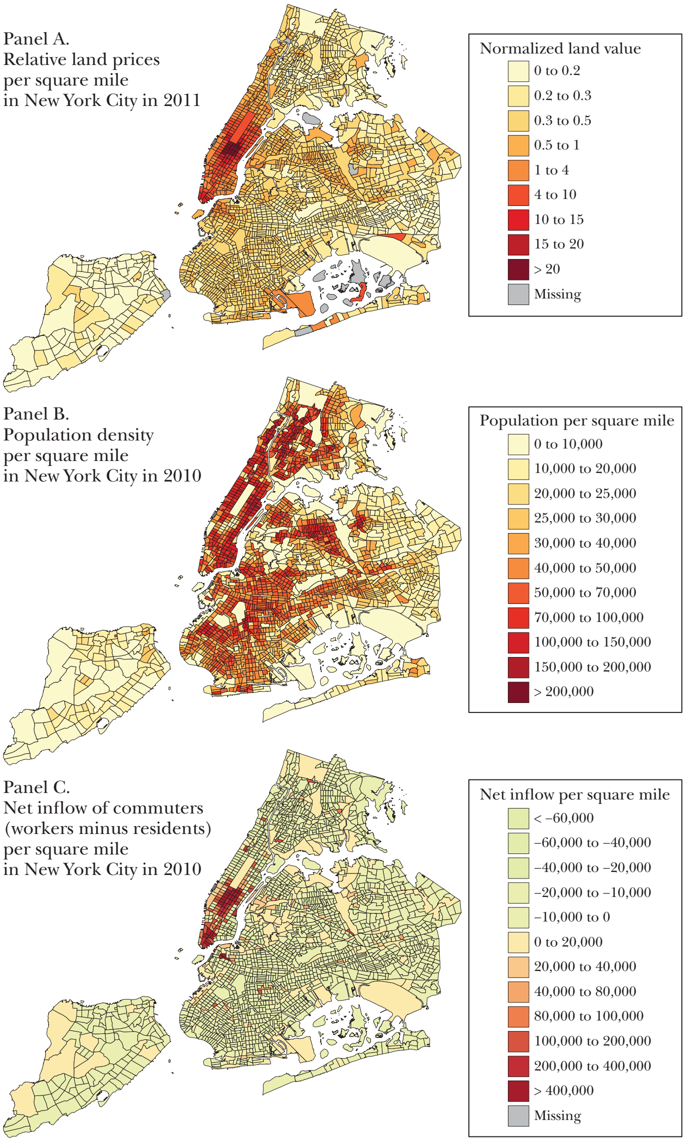
The three panels of Figure 1 illustrate this within-city variation by using census tract data across the five boroughs of New York City. The borough of Staten Island is in the lower left of the figure. The land directly east (across the Hudson River) includes Brooklyn to the west and Queens to the east. The island further north (between the Hudson River and the East River) includes Manhattan to the south and the Bronx to the north. Census tracts are intended to include about 4,000 people, although for localized reasons they can be half or twice that size.

■ *Stephen J. Redding is Professor in Economics, Princeton University, Princeton, New Jersey. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts, and a Research Fellow, Centre for Economic Policy Research, London, United Kingdom. His email address is reddings@princeton.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.75.

Figure 1

Economic Geography of New York City



Panel A.
Relative land prices
per square mile
in New York City in 2011

Panel B.
Population density
per square mile
in New York City in 2010

Panel C.
Net inflow of commuters
(workers minus residents)
per square mile
in New York City in 2010

Source: Panel A: Primary Land Use Tax Lot Output, NYC Department of City Planning. Panel B: US population census. Panel C: Data on commuting flows from the LEHD Origin-Destination Employment Statistics (LODES).

Note: Panel A: Assessed land values per square mile in 2011, normalized by the mean across census tracts in New York City, including the five boroughs of Bronx, Brooklyn, Manhattan, Queens, and Staten Island. A value above one corresponds to an above average land value per square mile. Land values are from property taxation assessments. Panel B: Population density per square mile in 2010 for each census tract in the five boroughs of New York City (the five boroughs of Bronx, Brooklyn, Manhattan, Queens, and Staten Island). Panel C: Net inflow of commuters (workers minus residents) per square mile. Workers equals employment by workplace, which is the sum of all inward commuting flows into a census tract from anywhere in the United States (including from the census tract itself). Residents equals employment by residence, which is the sum of all outward commuting flows from a census tract to anywhere in the United States (including to the census tract itself). Negative values represent net exports of commuters and positive values correspond to net imports of commuters.

Panel A shows the variation in land prices across New York City, measured using the assessed tax value of the land per square meter in 2011.[1] Midtown Manhattan has by far the highest land prices, with a smaller secondary peak in downtown Manhattan, and an area of lower land values in between. Towards the bottom of Manhattan, the Lower East Side has noticeably lower land prices than other nearby neighborhoods. Across the East River, central Brooklyn is the site of another smaller peak in land prices. Finally, the areas bordering Central Park in Manhattan are relatively more expensive than those further away from the park. The higher land prices in some neighborhoods relative to others can be explained in terms of the demand for either commercial or residential land use. In 2010, the top 10 percent of census tracts in New York City with the highest land value per square mile accounted for 60 percent of total land value and 64 percent of employment, but only 14 percent of population and 7 percent of land area.[2]

Panel B of Figure 1 displays the number of residents per square mile in 2010 for each census tract in New York City. Manhattan is the most densely-populated county in the United States: In 2010, 1,518,500 people lived in an area of 22.8 square miles, with a population density of 66,579 people per square mile. In contrast, Staten Island is relatively sparsely populated, with a population density an order of magnitude smaller at 7,923 people per square mile.[3] Even within Manhattan, population density displays dramatic variation, with relatively low densities in the commercial districts of midtown and downtown, and relatively high densities in the residential suburbs of the Upper West and Upper East Sides.

Panel C shows net imports of commuters (workers minus residents) per square mile in 2010 for each census tract in New York City. Workers corresponds to employment by workplace, which is the sum of all in-commuting flows to a census tract (including from the census tract itself). Residents equals employment by residence, which is the sum of all out-commuting flows from a census tract (including to the census tract itself). A positive value implies that a census tract is a net importer of commuters, while a negative value implies that it is a net exporter of commuters. Areas with high productivity relative to amenities can specialize as workplaces, while those with high amenities relative to productivity can specialize as residences. The result is a rich internal structure of economic activity within cities.

Location specialization is even more dramatic by this measure. The two land price peaks of midtown and downtown Manhattan are highly-specialized commercial

[1] Similar patterns are observed using land prices estimated from property transactions data. For evidence from land prices estimated using property transactions data, see for example Barr, Smith, and Kulkarni (2018) and Haughwout, Orr, and Bedoll (2008).

[2] These authors' estimates are from the Primary Land Use Tax Lot Output, NYC Department of City Planning, the US population census, and LEHD Origin-Destination Employment Statistics (LODES).

[3] Even so, Staten Island is densely-populated relative to many rural locations in the United States, with the state of Wyoming having a population density in 2010 of 5.6 people per square mile.

districts, with net imports of commuters greater than 400,000 per square mile.[4] Although Manhattan as a whole is a net importer of commuters, we find that parts of the island specialize as residences, with high exports of commuters per square mile in some areas on the Upper West and Upper East Side. Additionally, Brooklyn and the Bronx also have some census tracts with high imports of commuters per square mile. This specialization also occurs at the intermediate scale of the five boroughs of the city: In the same year, Manhattan alone was a net importer of 1.4 million workers from the rest of New York City and the tri-state area (with an inflow of 1.6 million and an outflow of 0.19 million). Moreover, this specialization occurs at the macro scale of New York City and its economic hinterland across the states of Connecticut, New Jersey, and New York: In 2010, New York City was a net importer of 0.5 million workers from the rest of this tri-state area (with an inflow of 0.97 million and an outflow of 0.46 million).

These rich patterns of the concentration of economic activity can be explained by a three-way interaction between natural advantages, agglomeration forces, and dispersion forces. In the context of New York City, the waterfront areas around the edge of Manhattan historically had *natural advantages* for production, port facilities, warehousing, and industrial processing, which could have long-lived effects on land values. In contrast, Central Park—in the midst of the densely-populated city—is an important natural amenity for consumption. Discussions of *agglomeration* typically feature externalities, such that an agent making a location decision does not take into account how that decision will affect the location decisions of other agents. These externalities can be either technological (say, knowledge spillovers between agents) or mediated through markets (say, demand for locally-traded goods and services). These agglomeration forces promote the concentration of economic activity, but are offset by *dispersion* forces. For example, when the price of local factors of production that are in inelastic supply are bid higher, such as land, incentives arise that shift production or residential activity to areas with lower land prices. More broadly, the concentration of economic activity can give rise to congestion or facilitate the spread of disease between people, both of which can act as dispersion forces.

The complexity of modeling these forces in spatial equilibrium has meant that the traditional theoretical literature on cities focused on stylized settings, such as a monocentric city with one central business district, a one-dimensional city on a line, or a perfectly symmetric circular city. I begin with a brief review of these earlier models, but as New York City and many other cities readily illustrate, such models cannot capture the rich internal variation in patterns of economic activity within real world cities, nor can they be easily used for detailed analysis of events or policies affecting a specific city.

[4] This figure is a density per land area, where census tracts can be much smaller than a square mile. Maximum and minimum net imports of commuters (without dividing by land area) are 190,292 and –7,390, respectively.

The main focus of this paper is to describe the recent development of quantitative urban models that connect directly with observed data on real world cities. These models allow for many locations within a city that can differ in productivity, amenities, land area, the supply of floor space, and transport connections. This heterogeneity across locations reflects the impact of natural advantages, agglomeration forces, and dispersion forces. Given the richness and flexibility of these quantitative urban models, they have been used to analyze a host of issues in urban economics: the strength of agglomeration forces, zoning and building regulations, the impact of transport infrastructure improvements, the sorting of heterogeneous groups of workers across space, and congestion pricing, among many others. These frameworks are sufficiently tractable that they permit a mathematical analysis of their properties, such as the conditions under which there is a unique equilibrium versus multiple equilibria in the model. In the presence of multiple equilibria, even small public policy interventions can have substantial effects, by shifting the location of economic activity between multiple equilibria. These theories also can be used to examine the effects of exogenous shocks affecting a city, like the division of Berlin by the Berlin Wall or the invention of a new mass-transit technology.

In the aftermath of the Covid-19 pandemic, these models can also provide insights as urban areas react to a range of issues; for example, concerns about disease, the practice of social distancing, a rise in remote work, changes in public transit systems, and further innovations in transport technology, such as ride-hailing and -sharing and autonomous vehicles.

Traditional Theoretical Models of Urban Economics

Traditional theoretical models in urban economics are focused on explaining stylized features of the data, such as the existence of a land price gradient, in which land prices are typically higher in the city center and on average decline with distance from the city center. Often these traditional models assume that economic activity is *monocentric*, in the sense that there is a well-defined central business district with a single peak of land prices. They may also consider a restrictive geography, such as identical locations along the real line or a perfectly symmetric circular city.

In the canonical model of internal city structure following Alonso (1964), Muth (1969), and Mills (1967), cities are monocentric by assumption. All employment is assumed to be concentrated in a central business district and workers face commuting costs in traveling to work. As workers living further from the city center face higher commuting costs, this must be compensated in equilibrium by a lower land rent further from the city center, in order for workers to be indifferent across locations. The geographical boundary of the city is determined by the return to land in its competing use of agricultural production. Therefore, a central prediction of these traditional theories is that land rents decline monotonically with distance from the city center, consistent with the observed property of the data that central locations on average command higher land prices than outlying areas.

Many cities with long histories of settlement (like London) are well approximated by this assumption of a monocentric pattern of economic activity. In contrast, other cities that developed more recently (like Los Angeles) are better described by a polycentric structure, in which there are multiple business districts spread throughout the metropolitan area. One polycentric structure is an "edge city," which consists of multiple concentrations of business, shopping, and entertainment outside a traditional downtown or central business district, often beside a major road in what had previously been a suburban residential or rural area.

To allow for the possibility of polycentricity, the assumption that all employment is concentrated in the city center can be relaxed to allow for an endogenous allocation of land between commercial and residential use throughout the city. Fujita and Ogawa (1982) consider the case of a one-dimensional city along the real line, while Lucas and Rossi-Hansberg (2002) analyze a perfectly symmetric circular city. In these frameworks, whether monocentric or polycentric patterns of economic activity emerge depends on the strength of agglomeration and dispersion forces. On the one hand, a nonmonocentric pattern of alternating areas of commercial and residential land use reduces commuting costs, because workers can typically live closer to their place of employment than in a monocentric structure. On the other hand, these alternating areas of commercial and residential land use reduce the concentration of employment, and hence diminish agglomeration economies relative to the monocentric case.

In summary, key insights from this theoretical literature are the role of the trade-off between agglomeration forces and commuting costs in generating urban rent gradients, and in determining whether these rent gradients are monocentric or polycentric.

Quantitative Urban Models

Although traditional models in urban economics explain certain features of the data, their simplifying assumptions of monocentricity or symmetry limit their usefulness for empirical work. These simplifying assumptions abstract from empirically relevant differences in natural advantage across locations, such as access to natural harbors or green parks. No city in the real world is perfectly monocentric or symmetric.

To address these limitations, recent quantitative urban models allow for empirically relevant differences in natural advantage while also incorporating agglomeration forces. These models are designed to connect directly to observed data on cities, which feature rich asymmetric patterns of economic activity—say, higher land prices in western than in eastern suburbs—and scattered clusters of employment and residents throughout a given city. Because these models connect directly to the observed data, they can be used to estimate the strength of agglomeration forces, or to undertake counterfactuals to predict the impact of realistic public policy interventions, such as the construction of a new subway line along a specific route within a given city.

We begin by describing a baseline quantitative version of the canonical urban model following Ahlfeldt et al. (2015), before discussing a number of extensions and generalizations.[5] Redding and Rossi-Hansberg (2017) review quantitative spatial models more broadly and Redding (2022a) surveys the wider literature on trade and geography.

Building Blocks

Consider a city, embedded in a wider economy. The city consists of a set of discrete "blocks" or census tracts. Each block has a supply of floor space that depends on its geographical land area and the density of development (the ratio of floor space to land area). Floor space can be used either commercially or residentially, or with some mixture of the two, a choice which will be influenced by zoning regulations.

The city is populated by workers, who are mobile between the city and the larger economy. Workers first decide whether to move to the city, and if so, they then consider each possible pair of residence and workplace blocks within the city. Workers have idiosyncratic preferences for living and working in different locations within the city. They consider all the personal, work-related, or amenity-related reasons for living in one place and working in another, and pick the residence-workplace pair that yields the highest utility. Commuting costs increase with the travel time between the worker's residence and workplace. Residential amenities depend on both natural advantages, such as leafy streets and scenic views, and agglomeration forces in the form of residential externalities, including positive externalities from nontraded goods and negative externalities from crime.

Because the model is focused on location choices within the city, it assumes away different kinds of final goods. Instead it assumes a single final good that is costlessly traded both within the city and with the wider economy, within perfectly competitive markets. This final good is produced using inputs of labor and commercial floor space according to a constant returns to scale technology. However, productivity of the final good can differ across locations within the city and depends both on natural advantages, such as access to natural water, and on agglomeration forces in the form of production externalities, which depend on employment density in surrounding locations (and are influenced by knowledge spillovers).

Economic Forces

We now discuss the three sets of economic forces that shape the equilibrium organization of economic activity within the city: productivity differences across locations, amenity differences across locations, and the transportation network.

High productivity in a location raises the marginal productivities of labor and land, which increases wages and the price of commercial floor space. In contrast, high amenities in a location raise the utility of living there, which attracts residents,

[5] An accompanying online Appendix provides a more detailed development of the model and a formal characterization of its theoretical properties.

and bids up the price of residential floor space. Transportation networks allow workers to separate where they live from where they work to take advantage of these differences in productivity and amenities. Through this separation of home and work, some locations specialize as workplaces (often but not always in central cities), while other locations specialize as residences (often but not always in outlying suburbs).

In locations with high productivity relative to amenities, the return to commercial land use exceeds the return to residential land use. Therefore, these locations specialize as workplaces, with higher employment than residents and net imports of commuters. In contrast, in locations with high amenities relative to productivity, the converse is true. The return to residential land use exceeds the return to commercial land use, such that these locations specialize as residences, with lower employment than residents and net exports of commuters. If a location has both positive employment and positive residents, either the return to commercial land use equals the return to residential land use, or zoning regulations sustain a wedge between the returns to commercial and residential land uses.

The resulting commuting flows between locations are typically assumed to satisfy a *gravity equation*, so-called because of the parallel with the Newtonian theory of gravity. According to this specification, the bilateral flow of commuters between a residence and workplace is decreasing in bilateral commuting costs, increasing in the attractiveness of the workplace (for example, as captured by its wage), and increasing in the attractiveness of the residence (for example, as shaped by its amenities). The attractiveness of each residence also depends on its overall commuting costs to all workplaces, often referred to as "multilateral resistance." Even if a specific workplace-residence pair has high bilateral commuting costs (high *bilateral resistance*), we may still observe substantial bilateral commuting flows if that residence has even higher commuting costs for all other workplaces (high *multilateral resistance*). This gravity equation specification is both theoretically tractable and provides a good approximation to observed commuting patterns (for example, as in Fotheringham and O'Kelly 1989; McDonald and McMillen 2010). Bilateral commuting costs depend on bilateral travel times, which can be computed using the observed transport network (say, underground and overground railway lines) and assumptions about the average speed of each mode of transport.[6]

The differences in productivity and amenities that induce location specialization and commuting reflect the combined impact of natural advantages and agglomeration forces. As employment concentrates in a location because of natural advantages for production, this raises employment density, which further increases productivity through production externalities, thereby magnifying the impact of differences in natural advantage. Similarly, most empirical studies find net residential externalities to be positive. Therefore, as residents concentrate in a location because of natural advantages for amenities, this further increases amenities

[6] A long line of research in transportation economics following McFadden (1974) models individuals' choice of transport mode.

through residential externalities, again magnifying the impact of differences in natural advantage.

These production and residential agglomeration forces are offset by dispersion forces from commercial and residential floor space use. As employment concentrates in a location, this bids up the price of commercial floor space, which raises firms' cost, and encourages the dispersion of employment to lower density locations. Similarly, as residents concentrate in a location, this bids up the price of residential floor space, which reduces worker utility and encourages the dispersion of residents to lower density locations. Although the overall demand for floor space in a location can be reduced by separating workplace and residence, this gives rise to commuting costs, which themselves provide another force for the dispersion of economic activity.

The strength of this dispersion force from a limited supply of floor space depends on assumptions about the supply elasticity for floor space. Floor space is typically assumed to be produced by a competitive construction sector that uses land and capital as inputs. Land itself is in perfectly inelastic supply. In contrast, for a city that is small relative to the wider economy, capital is in perfectly elastic supply at an exogenous cost of capital. As a result, the smaller is the share of land in construction costs, the larger is the supply elasticity for floor space, and the weaker are dispersion forces.

If production and residential agglomeration forces are sufficiently strong relative to these dispersion forces, the spatial organization of economic activity within the city can be subject to multiple equilibria. If employment is expected to concentrate in some locations and residents are expected to concentrate in other locations, this itself generates differences in productivity and amenities that can support such specialization as an equilibrium outcome. At small spatial scales within cities where natural advantages are similar, it is particularly plausible that the location of economic activity could be subject to such multiple equilibria, especially for individual economic functions (for example, whether shoe stores are clustered on one street rather than a neighboring street).

The general equilibrium of the model satisfies the following equilibrium conditions: (1) cost minimization and zero profits in production; (2) utility maximization and population mobility; (3) cost minimization and zero profits in construction; (4) demand for commercial floor space equals the supply of commercial floor space; (5) demand for residential floor space equals the supply of residential floor space; (6) no-arbitrage between alternative uses of floor space, such that locations are either specialized as workplaces, specialized as residences, or incompletely specialized, depending on the relative values of the prices of commercial and residential floor space; (7) employment in each workplace is equal to the number of residents choosing to commute to that workplace.

From the zero-profit condition in production, firms must make zero profits in all locations with positive employment. Therefore, high productivity in a location must be offset in equilibrium by some combination of higher wages and/or a higher price of commercial floor space. Given data on wages and the price of commercial floor space, and an assumption about the production technology, it follows that one can use this zero-profit condition to back out the productivity required for the observed data to be an equilibrium of the model.

From the population-mobility condition, residents must be indifferent across all locations with positive residents. Hence, high amenities in a location must be offset in equilibrium by some combination of lower expected income net of commuting costs and/or a higher price of residential floor space. Given data on wages, travel times, and the price of residential floor space and an assumption about the utility function, it also follows that one can use this population mobility condition to back out the amenities required for the observed data to be an equilibrium of the model.

Through these differences in productivity and amenities across locations, quantitative urban models are able to rationalize the rich polycentric and asymmetric patterns of economic activity observed in real-world cities. Given the values of productivity and amenities backed out from the observed data, quantitative urban models can be used to structurally estimate the role of agglomeration forces in determining these variables. Before illustrating this, we next provide further intuition for the determination of equilibrium in these models.

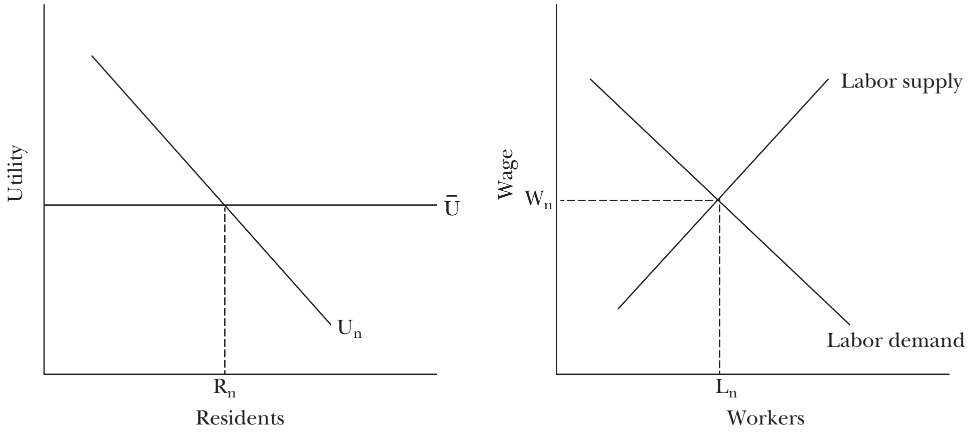
Equilibrium City Structure

We can gain a better sense of how internal city structure is determined at an intuitive level by looking at residential and workplace choices in partial equilibrium, where these two sets of decisions are then linked in general equilibrium. This partial equilibrium analysis connects closely with conventional models of labor demand and supply.

In the left-hand panel of Figure 2, we illustrate the determination of the number of residents (R_n) who choose to live in a certain location in the city. The horizontal line shows the reservation level of utility in the wider economy ($\bar{U}$). The downward-sloping line shows expected utility from residence n (U_n). Expected utility is decreasing in the number of residents for two reasons. First, as we increase the number of residents for a given supply of residential floor space, this bids up the price for residential floor space, which reduces expected utility. Second, as we increase the number of residents in a given location, we attract workers with lower idiosyncratic preferences for that location, which reduces expected utility through a composition or batting-average effect.

The equilibrium number of residents is determined by the intersection of the reservation level of utility in the wider economy ($\bar{U}$) and the expected utility from living in location n. The position of the expected utility curve (U_n) depends on amenities in location n and expected income net of commuting costs from access to workplaces from that location. An increase in amenities in location n (as an example, perhaps from improved access to green spaces) shifts outwards the expected utility curve (U_n), which increases the number of residents in location n (R_n). Similarly, an improvement in location n's connections to the transport network increases expected income net of commuting costs from that location, which shifts outwards

Figure 2
Residence and Workplace Choices



Source: Author's analysis.
Note: Left-hand panel shows the partial equilibrium determination of the number of residents in location n (R_n) by the reservation utility in the wider economy ($\bar{U}$) and the expected utility of living in that location (U_n). Right-hand panel shows the partial equilibrium determination of workers in location n (L_n) by labor demand and supply as a function of the wage (w_n) in that location. See the online Appendix for a formal derivation of this diagram.

the expected utility curve (U_n), and increases the number of residents in location n (R_n).

In the right-hand panel of Figure 4, we illustrate the determination of the number of workers in each workplace (L_n). The downward-sloping line shows labor demand in workplace n, as determined by the equality between the wage and the value marginal product of labor. An increase in the number of workers employed in a location leads to a decrease in the wage because of diminishing marginal physical productivity of labor in the production technology. The upward-sloping line shows labor supply for workplace n, as determined by worker choices of residence and workplace. In order to increase labor supply, firms must offer a higher wage in order to attract workers with lower idiosyncratic preferences for that workplace.

The position of the labor demand curve depends on productivity in location n, while the position of the labor supply curve depends on the transport network and access to commuters from surrounding locations. An increase in productivity in location n (as an example, perhaps from improved access to natural water) raises the marginal product of labor, which shifts outwards the labor demand curve, and increases employment in location n (L_n). An improvement in location n's connections to the transport network increases the supply of commuters at a given wage, which shifts outwards the labor supply curve and increases employment in location n (L_n).

From the left-hand panel, a location has zero residents, and hence completely specializes as a workplace, if expected utility (U_n) always lies below the reservation level of utility in the wider economy ($\bar{U}$) for all positive values of residents (R_n). From the right panel, a location has zero employment, and hence completely specializes as a residence, if the labor demand curves lies below the labor supply curve for all positive values of workers (L_n). More generally, a location can either be a net importer of commuters if employment exceeds residents ($L_n > R_n$), or a net exporter of commuters if employment falls short of residents ($L_n < R_n$).

Although Figure 4 provides useful intuition, it is important to keep in mind that it only provides a partial equilibrium analysis and does not capture all general equilibrium relationships in the model. First, this figure focuses on the commuter market for residents and workers, but the position of the expected utility and labor demand curves is also influenced by the land market. Second, the expected utility curve in the left panel (U_n) is jointly determined with the labor supply curve in the right panel within a given location, because residents can work locally. Third, the expected utility curve in the left panel (U_n) for one location is jointly determined with the labor supply curve for other locations, because residents in one location commute to work in other locations. Therefore, an increase in amenities in surrounding locations, which raises the number of surrounding residents, increases a location's own supply of labor. Fourth, amenities and productivity depend on surrounding concentrations of residents and workers, respectively, through agglomeration forces. In Figure 4, we have assumed that these agglomeration forces are not too strong, such that expected utility (U_n) is downward-sloping in a location's own residents (R_n), and labor demand is downward-sloping in a location's own workers (L_n).

Extensions and Generalizations

Although we have considered a relatively parsimonious quantitative urban model, the tractability of these frameworks lends itself to a large number of extensions and generalizations, which can be used to address a range of public policy issues.

First, this class of models can accommodate nontraded goods, as in Heblich, Redding, and Sturm (2020). Second, the models can accommodate other reasons for travel apart from commuting, such as consumption trips, as in Miyauchi, Nakajima, and Redding (2022). Third, they can allow for multiple final goods with costly trade and technology differences, as in Eaton and Kortum (2002) and Redding (2016). Fourth, they can encompass final goods that are differentiated by origin and costly trade, as in Armington (1969), Allen and Arkolakis (2014), and Allen, Arkolakis, and Li (2017). Fifth, these quantitative urban models can encapsulate horizontally-differentiated firm varieties with costly trade, as in Helpman (1998), Redding and Sturm (2008), and Monte, Redding, and Rossi-Hansberg (2018).

Sixth, they can be used to quantify the impact of zoning regulations on internal city structure, as in Allen, Arkolakis, and Li (2017). Seventh, they can be used as a platform for evaluating neighborhood development programs, as in the analysis of the redevelopment of Detroit in Owens, Rossi-Hansberg, and Sarte (2020). Eighth, this kind of model can incorporate forward-looking investments in capital accumulation, as in Kleinman, Liu, and Redding (2023). Ninth, it can allow for multiple groups of workers that are heterogeneous, as in Redding and Sturm (2016) and Tsivanidis (2018). Tenth, whereas travel time was treated as exogenous and independent of commuting flows above, congestion can be introduced, as in Allen and Arkolakis (2022). Which of these specifications is most useful for empirical work depends on the data available and the public policy issue of interest.

Recent events have drawn attention to a range of public policy issues that can be addressed using quantitative urban models. The outbreak of the Covid-19 pandemic reminded us that disease contagion has been a powerful dispersion force throughout human history (for example, as discussed in Glaeser and Cutler 2021). New technologies and forms of managerial organization that allow remote or hybrid working can be interpreted as reductions in commuting costs, as workers no longer need to travel from their home to their workplace or are only required to do so for a smaller number of days each week (see the discussion in Barrero, Bloom, and Davis 2021). Similarly, autonomous vehicles can be interpreted as another technological innovation that reduces commuting costs. To the extent that workers no longer need to pay attention while driving, this will free up additional time for work or leisure. If an active ride-hailing market develops for autonomous vehicles, this may also free up substantial areas of land in urban areas that are currently used for parking private vehicles. In the empirical applications below, we provide another example from history of how a technological innovation (the invention of the steam passenger railway) reduced commuting costs and reshaped patterns of specialization by residence and workplace within urban areas.

Application 1: The Division of Berlin

Throughout the long literature on economic geography and urban economics, it has been empirically challenging to distinguish agglomeration and dispersion forces from variation in natural advantages. After all, high land prices and levels of economic activity in a group of neighboring locations are consistent with strong agglomeration forces, but equally consistent with shared amenities that make these locations attractive places to live (like leafy streets and scenic views) or common natural advantages that make these locations attractive for production (like access to natural water).[7] To disentangle these two alternative explanations for location choices, one requires a source of exogenous variation in the surrounding concentration of economic activity. Ahlfeldt et al. (2015) uses the division of Berlin in the aftermath of the Second World War and its reunification following the fall of the Iron Curtain as such a source of exogenous variation.

[7] This is an example of the broader challenge in the social sciences of distinguishing spillovers from correlated individual effects, as discussed in Manski (1995).

The Qualitative Story

A protocol signed in London in September 1944 near the end of World War II designated separate occupation sectors in Berlin, Germany, for the American, British and Soviet armies. The boundaries between these occupation sectors were chosen based on pre-war administrative districts that had little prior significance, such that the three sectors were of roughly equal population, with the Americans and British in the West, and the Soviets in the East. Later a French sector was created from part of the British sector. The original plan was for Berlin to be administered jointly by a central committee ("Kommandatura"). However, following the onset of the Cold War, East and West Germany were founded as separate states, and separate city governments emerged in East and West Berlin in 1949. For a while travel between the different sectors of Berlin remained possible, until, to stop civilians leaving for West Germany, the East German authorities constructed the Berlin Wall in 1961.

Ahlfeldt et al. (2015) provides evidence that Berlin's land price gradient in 1936 was approximately monocentric, with the highest values concentrated in the prewar central business district in the neighborhood of Mitte, with concentric rings of progressively lower land prices in the surrounding areas. However, Mitte was east of the future line of the Berlin Wall and thus was cut off when the wall was built. If one looks only at the areas of Berlin in 1936 that were going to become part of the future West Berlin, the two parts of the future West Berlin with the highest land prices in 1936 were an area just west of the prewar central business district and the future line of the Berlin Wall, and the Kudamm ("Kurfürstendamm") further west, which had developed into a fashionable shopping area in the decades leading up to World War II.

By 1986, looking at West Berlin following division, we find that the first prewar land price peak just west of the prewar central business district is entirely eliminated. This area ceased to be an important center of commercial and retail activity. Instead, the second prewar price peak in what had been the secondary area of the Kudamm develops into West Berlin's central business district during the period of division.

By 2006, after the reunification of Berlin, the prewar central business district that had been in the former East Berlin reemerges as a land price peak, as does the area just west of this prewar central business district and the former line of the Berlin Wall, which is now again a concentration of office and retail development.

These patterns are consistent with the qualitative predictions of the model developed above. Following division, the biggest declines in land prices are observed in the parts of West Berlin closest to the pre–World War II city's central business district. These parts of West Berlin experience the greatest reductions in access to production agglomeration forces, residential agglomeration forces, supplies of commuters, and employment opportunities from the areas of the prewar city that became East Berlin. There is also little evidence of an impact on land prices along other sections of the Berlin Wall following division. This pattern of results supports the idea that it is not proximity to the Berlin Wall per se that matters, but rather the loss of access to nearby concentrations of employment and residents in East Berlin.

These observed changes in the land price gradient are accompanied by a similar reorientation of employment and residents within West Berlin.

Quantitative Evidence

To examine whether the quantitative urban model developed above can account for the observed changes in the spatial distribution of land prices, employment, and residents, Ahlfeldt et al. (2015) estimate the structural model's parameters. Using a given set of parameters, the structure of the model can be used to solve for the unobserved values of natural advantages for production, natural advantages for amenities, and the density of development (as measured by the ratio of floor space to land area). With this estimation procedure, the model exactly rationalizes the observed data on land prices, employment, and residents in each year before and after division and reunification as an equilibrium outcome.

The model's parameters are estimated using the identifying assumption that changes in natural advantages for production and amenities in each city block are uncorrelated with the change in the surrounding concentration of economic activity induced by Berlin's division and reunification. Because the city's division stemmed from military considerations during World War II and its reunification originated in the wider collapse of Soviet communism, the resulting changes in the surrounding concentration of economic activity are plausibly exogenous to changes in natural advantages in individual city blocks. In particular, these changes in natural advantages in West Berlin are assumed to be orthogonal to indicator variables for distance of grid cells to the prewar central business district. This identifying assumption requires that the systematic change in the gradient of economic activity in West Berlin relative to the prewar central business district following the city's division is explained by the mechanisms of the model—that is, by the changes in commuting access and production and residential agglomeration forces—rather than by systematic changes in natural advantages for production and amenities. The analysis focuses on West Berlin, because it remained a market economy, and hence one would expect the mechanisms in the model to apply. In contrast, allocations in East Berlin during the period of division were determined by central planning, which is unlikely to mimic market forces.

The parameters are estimated for both division and reunification separately, and then by pooling all of the data together. All three specifications yield a similar pattern of estimated coefficients, with evidence of substantial agglomeration forces from production and residential externalities. In the specification pooling both sources of variation, the estimated elasticity of productivity with respect to travel-time-weighted employment density is 0.07, while the estimated elasticity of amenities with respect to travel-time-weighted residents' density is 0.15. These agglomeration forces are highly localized. The estimates imply that both production and residential externalities fall to close to zero after around ten minutes of travel time, which corresponds to around 0.83 kilometers by foot (at an average speed of five kilometers per hour) and about four kilometers by underground and suburban railway (at an average speed of 25 kilometers per hour).

Other Evidence

These parameter estimates from Berlin's division and reunification are broadly consistent with the findings of other empirical research. The estimate of the elasticity of productivity with respect to production externalities of 0.07 is towards the high end of the 3–8 percent range from the survey by Rosenthal and Strange (2004), but less than the elasticities from some quasi-experimental studies (for example, Greenstone, Hornbeck, and Moretti 2010; Kline and Moretti 2014).

The finding of highly localized production externalities is also consistent with other research using within-city data. Using data on the location of advertising agencies in Manhattan, Arzaghi and Henderson (2008) find little evidence of knowledge spillovers beyond 500 meters straight-line distance. In comparison, a straight-line distance of 450–550 meters in Berlin corresponds to around nine minutes of travel time, after which production externalities are estimated to have declined to around 4 percent.

Finally, the finding of substantial residential externalities is in line with recent empirical findings that urban amenities are endogenous to the surrounding concentration of economic activity (Glaeser, Kolko, and Saiz 2001; Diamond 2016; Almagro and Domínguez-Lino 2022). Similarly, using data on an urban revitalization program in Richmond, Virginia, Rossi-Hansberg, Sarte, and Owens (2010) also find residential externalities are highly localized, with housing externalities falling by approximately one-half every 1,000 feet.

Taking the empirical findings of this section together, the quantitative urban model developed above is able to rationalize the rich patterns of spatial variation in land prices, employment, and residents observed in the data. Furthermore, for the estimated parameter values, the model is quantitatively successful in predicting the change in the internal city structure in response to the large-scale shock of Berlin's division and reunification.

Application 2: The Nineteenth-Century Steam Railway Revolution in London

The dense concentrations of economic activity observed in modern metropolitan areas involve transporting millions of people each day between their home and place of work. For example, the London Underground today handles around 3.5 million passenger journeys a day, and its trains and its trains travel around 76 million kilometers (about 47 million miles) each year. What is the role of London's transport network in sustaining its dense concentrations of economic activity? Heblich, Redding, and Sturm (2020) use the mid-nineteenth-century invention of steam railways as a natural experiment to explore this question. The key idea is that steam railways dramatically reduced travel time for a given distance, thereby lowering commuting costs, and permitting the first large-scale separation of workplace and residence.

Greater London provides an attractive empirical setting for this analysis, because of the availability of spatially-disaggregated data on economic activity over a long time horizon from 1801 to 1921, before and after this transport innovation. Data are available for a number of different geographical definitions of London. Greater London, as defined by the boundaries of the modern Greater London Authority (GLA), includes a 1921 population of 7.39 million and an area of 1,595 square kilometers. The historical County of London has a 1921 population of 4.48 million and an area of 314 square kilometers. The City of London has a 1921 population of 13,709 and an area of about three square kilometers, and its boundaries correspond approximately to the Roman city wall.

The Qualitative Story

At the beginning of the nineteenth century, the most common mode of transport in London was walking, with average travel speeds in good road conditions of around three miles per hour. When the horse omnibus started in London in the 1820s, average travel reached perhaps six miles per hour. However, the opening of the London and Greenwich railway in 1836 as the first steam railway to be built specifically for passengers transformed the relationship between travel time and distance, with average travel speeds of around 21 miles per hour.

The availability of the steam passenger railway was followed by a large-scale change in the organization of economic activity within Greater London. In the first half of the nineteenth century, population in the City of London was relatively constant (at around 130,000), while population in Greater London grew substantially (from 1.14 million to 2.69 million). From 1851 onwards, shortly after the first steam passenger railways, population in the City of London falls sharply by around 90 percent to 13,709 in 1921. In contrast, the population of Greater London as a whole continues to grow rapidly from 2.69 million in 1851 to 7.39 million in 1921.

In the City of London, we observe the emergence of the first large-scale separation between the night population (where people sleep) and the day population (where they work). In the opening decades of the nineteenth century, the night and day populations for the City of London are relatively similar at about 150,000. But in the decades following the first steam passenger railways, in the City of London day censuses for 1866, 1881, 1891 and 1911, the sharp decline in night population is combined with a steep rise in day population. By 1911, the day population of the City of London was approaching 400,000, while the night population had fallen to only 10,000. This pattern of empirical results is consistent with the idea that the reduction in commuting costs from this new transport technology allowed the City of London to specialize as a workplace (importing commuters), while the surrounding suburbs specialized as residences (exporting commuters).

Quantitative Evidence

To rationalize these empirical findings, Heblich, Redding, and Sturm (2020) develop an estimation procedure that illustrates how quantitative urban models can be used to undertake counterfactuals for transport infrastructure improvements or other public policy interventions. Given data on economic activity in an initial observed equilibrium and estimates of the changes in travel times from a transport improvement, these models can be solved for the predicted change in the spatial organization of economic activity. Using these predictions, the economic benefit from the transport improvement can then be compared to estimates of its construction costs.

As a first step in implementing this procedure, the relationship between commuting costs and travel times is estimated using data on bilateral commuting flows and the observed transport network in London in the year 1921. This transport network includes overground and underground railways, buses and trams, and walking, since commuting by private car was negligible in London in 1921. Because the placement of transport infrastructure is potentially endogenous, this estimation uses an instrumental variable for travel time using the transport network in the form of bilateral geographical distance between locations. Given these estimates and the observed evolution of the transport network over time, predicted changes in commuting costs from the expansion of the railway network can be calculated.

Armed with these estimates of changes in commuting costs, observed data on bilateral commuting flows for 1921, and data on property values and employment by residence in earlier decades, the model can be solved for predicted employment by workplace and commuting flows back to the beginning of the nineteenth century. An advantage of using these historical data on property values and employment by residence is that the values of these variables in earlier decades can be used to control for other factors that changed over time in addition to the transport network, such as productivity or amenities.

The model successfully captures the observed sharp divergence between the night and day populations in the City of London from the mid-nineteenth century onwards. As the improvement in transport technology reduces commuting costs, workers become able to separate their residence and workplace to take advantage of high wages in locations with high productivity relative to amenities (so that these locations specialize as workplaces) and the lower cost of living in locations with high amenities relative to productivity (so that these locations specialize as residences). If productivity and amenities depend on the density of workers and residents, respectively, through agglomeration forces, this concentration of employment in the center and dispersion of population to the suburbs further magnifies these differences in productivity and amenities across locations.

Although the City of London experiences by far the largest absolute increase in employment, the highest percentage rates of growth of employment (and population) occur in the suburbs, as these areas are transformed from villages and open fields to developed land. As a result, the gradient of employment density with respect to distance from the center of the City of London declines between 1831 and 1921, and the share of the 13 boroughs within five kilometers of the Guildhall in total workplace employment in Greater London falls from around 68 percent in 1831 to about 48 percent in 1921. This pattern of results is in line with a long line of empirical research that finds evidence of employment (and population) decentralization in response to transport improvements, as reviewed in Redding and Turner (2015) and Redding (2022b). These findings suggest that present technological changes, such as innovations in remote working and autonomous vehicles, have the potential to further decentralize economic activity.

As a specification check, the model's predictions for commuting flows are compared to historical data from the personnel ledgers of Henry Poole Tailors, a high-end bespoke tailoring firm, which was founded in 1802. The firm collected data on workers' residential addresses at the time they were first hired, thus allowing an estimate of commuting distances to the firm. There are of course a number of possible reasons why the pattern of employee commutes to a particular firm at a specific site might differ from the model's predictions. Nevertheless, the model is remarkably successful in capturing the change in the distribution of commuting distances between these time periods. In the opening decades of the railway age in the 1850s and 1860s, most workers in Westminster in both the model and data lived within five kilometers of their workplace. By the turn of the twentieth century, commuting distances up to 20 kilometers are observed in both the model and data.

Evaluation of Transport Infrastructure Investments

The estimated model also can be used to evaluate the economic benefits of the construction of London's railway network, holding constant all other factors, such as productivity and amenities. In this analysis, the impact of the railway network on worker utility depends on assumptions about labor mobility and land ownership. In particular, suppose that the economy consists of workers who own only labor and landlords who own only land, and assume that workers are perfectly mobile between London and the wider economy at an unchanged reservation level of utility. In this case, as the construction of London's railway network reduces commuting costs and raises expected worker utility, it attracts a population inflow, which bids up the price of land, until expected worker utility in London in the new equilibrium is equal to the unchanged reservation level of utility in the wider economy. Under these assumptions, all economic benefits from the construction of London's railway network accrue to landlords through a higher price of land. More generally, if labor is imperfectly mobile between London and the wider economy, the economic benefits from the railway network are enjoyed by both workers and landlords.

Under a range of different assumptions about labor mobility, the economic benefits from the construction of London's railway network are found to exceed historical estimates of its construction costs based on the capital issued by railway companies. The ratio of benefits to costs is substantially larger once production and residential agglomeration forces are taken into account. In the presence of these forces, the population inflow induced by the reduction in commuting costs induces endogenous increases in productivity and amenities. Similarly, the ratio of benefits to costs is enhanced by taking into account complementary investments in buildings and structures. The reason is that the resulting population inflow raises the demand for commercial and residential floor space, which leads to an endogenous increase in the supply of floor space from the construction sector. An important takeaway for these findings is the need to take into account agglomeration forces and complementary investments in buildings and structures in conventional cost-benefit analyses of transport infrastructure investments.

Looking beyond this empirical application, policymakers are often interested in comparing alternative possible transport investments, such as which links in a railway or highway network to improve. To develop a framework to address this question, Allen and Arkolakis (2022) embed a specification of endogenous route choice in a quantitative spatial model. In their approach, individuals consider travel costs and choose the least-cost route. A key implication of this framework is that the welfare effects of a small improvement in a transport link are equal to the percentage cost saving multiplied by the initial value of travel along that link.[8] Barwick et al. (2020) use an approach along these lines for an analysis of China's High Speed Rail Network, while Gupta, Van Nieuwerburgh, and Kontokosta (2022) provide evidence on the quantitative impact of the latest expansion to New York's subway network, the Second Avenue Subway.

More generally, Fajgelbaum and Schaal (2020) develop a framework for characterizing optimal transport networks in spatial equilibrium. This characterization is challenging, because the problem is high dimensional. However, they show that the problem of finding the optimal transport network can be transformed into the problem of finding the optimal flow in a network, which has been studied in the operations research literature. While this approach has so far been applied to trade in goods between cities, incorporating commuting within cities is an exciting avenue for further research.

Conclusion

Real-world cities feature complex internal structures, with a rich specialization by residential and commercial land use and an intricate division of labor. The real-world cities in which people live often exhibit dramatic changes in land prices and land use, both across neighborhoods and across blocks within neighborhoods. A key breakthrough in recent research has been the development of quantitative urban models that are able to rationalize and to explore these observed features of the data. These frameworks can accommodate many locations that differ in productivity, amenities, land area, the supply of floor space, and transport connections. Nevertheless, these models remain tractable and amenable to theoretical analysis with a manageable number of parameters to be estimated.

[8] Although this result is derived for particular functional forms, this implication is closely related to the celebrated result of Hulten (1978) that a sufficient statistic for the welfare effect of a small productivity shock in an efficient economy can be summarized by the appropriate Domar weight.

One key insight from these quantitative urban models is that the observed concentration of economic activity within cities cannot be explained by natural advantages alone, but instead requires substantial agglomeration forces. Another insight is the role of advanced transport networks in sustaining dense concentrations of economic activity in modern metropolitan areas.

An exciting area for further research is distinguishing between different underlying economic mechanisms for agglomeration. Although the quantitative urban model outlined above allows for agglomeration forces, these agglomeration forces are assumed to be reduced-form functions of travel-time-weighted employment density for production externalities and travel-time-weighted residents density for residential externalities.

However, following Marshall (1920), three main sets of forces for agglomeration are traditionally distinguished, which reflect the costs of moving goods, people, and ideas. First, firms may locate near suppliers or customers in order to save on transportation costs. Second, workers and firms may cluster together to pool specialized skills. Third, physical proximity may facilitate knowledge spillovers, as (in Marshall's words) "the mysteries of the trade become no mystery, but are, as it were, in the air." Another line of research dating back to Smith (1776) emphasizes a greater division of labor in larger markets, as examined empirically in Duranton and Jayet (2011). More recently, Duranton and Puga (2004) distinguish between sharing, matching, and learning as alternative mechanisms for the agglomeration of economic activity.

Although these mechanisms are well understood conceptually, there is relatively little evidence on their empirical importance, with a few exceptions such as Ellison, Glaeser, and Kerr (2010). Over time, the nature of economic activity undertaken within cities has changed dramatically, from the marketplaces and ports of pre-industrial Europe, through the centers of manufacturing of the industrial revolution, through the concentrations of office space of the mid-twentieth century, to an increasing focus on the consumption of nontraded goods and services in the twenty-first century. Using the verbs from occupational descriptions, Michaels, Rauch, and Redding (2018) quantify the change in the tasks undertaken by workers in cities over time. Whereas the tasks most concentrated in cities in 1880 involved the manipulation of the physical world, such as "Thread and Sew," those most concentrated in cities in 2000 involve human interaction, such as "Advise and Confer."

Given these large-scale changes in the types of economic activities performed in urban areas over time, it is plausible to think that the nature and scope of agglomeration economies could have evolved as well. Consistent with this idea, Autor (2019) finds substantial changes in the urban wage premium for workers with different levels of skills over time. At the beginning of his sample period in the 1970s, average wages were sharply increasing in population density for both low-skill workers (high-school or less) and high-skill workers (some college or greater). By the end of the sample period in 2015, this wage premium to population density had increased for high-skill workers but almost disappeared for low-skill workers.

Looking ahead, the wealth of newly-available sources of Geographical Information Systems (GIS) data promises to offer new opportunities to distinguish between different mechanisms for agglomeration, including ride-hailing data (from firms like Uber and Lyft), smartphone data with Global Positioning System (GPS) information, firm-to-firm data from sales (or value-added tax) tax records, credit card data with consumer and firm location, barcode scanner data with consumer and firm location, public transportation commuting data, work-from-home data, and satellite imaging data.

Over the centuries, cities have changed drastically—from marketplaces, to the locus of manufacturing industry, to clusters of office and retail development, and to centers of consumption. But as long as there are benefits to reduced costs of moving people, goods, and ideas, cities in some form are likely to thrive and prosper.

■ *I am grateful to Princeton University for research support. Thanks to Nan Xiang for excellent research assistance. I am very grateful to the editors for their extremely helpful comments and suggestions. I also would like to thank Gabriel Ahlfeldt, Treb Allen, Jason Barr, Dave Donaldson and Clifford Winston for helpful comments and suggestions. Any opinions, errors and omissions are the author's responsibility alone.*

References

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. "The Economics of Density: Evidence from the Berlin Wall." *Econometrica* 83 (6): 2127–89.

Allen, Treb, and Costas Arkolakis. 2014. "Trade and the Topography of the Spatial Economy." *Quarterly Journal of Economics* 129 (3): 1085–1140.

Allen, Treb, and Costas Arkolakis. 2022. "The Welfare Effects of Transportation Infrastructure Improvements." *Review of Economic Studies* 89 (6): 2911–57.

Allen, Treb, Costas Arkolakis, and Xiangliang Li. 2017. "Optimal City Structure." Unpublished.

Almagro, Milena, and Tomás Domínguez-Iino. 2022. "Location Sorting and Endogenous Amenities: Evidence from Amsterdam." University of Chicago Becker Friedman Institute for Economics Working Paper 2022-162.

Alonso, William. 1964. *Location and Land Use.* Cambridge MA: Harvard University Press.

Armington, Paul S. 1969. "A Theory of Demand for Products Distinguished by Place of Production." *International Monetary Fund Staff Papers* 16 (1): 159–78.

Arzaghi, Mohammad, and J. Vernon Henderson. 2008. "Networking Off Madison Avenue." *Review of Economic Studies* 75 (4): 1011–38.

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

Barr, Jason, Fred H. Smith, and Sayali J. Kulkarni. 2018. "What's Manhattan Worth? A Land Values Index from 1950 to 2014." *Regional Science and Urban Economics* 70: 1–19.

Barrero, Jose Maria, Nicholas Bloom, and Steven J. Davis. 2021. "Why Working from Home Will Stick." NBER Working Paper 28731.

Barwick, Panle Jia, Dave Donaldson, Shanjun Li, and Yatang Lin. 2020. "The Welfare Effects of Passenger

Transportation Infrastructure: Evidence from China." Unpublished.

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

Duranton, Gilles, and Hubert Jayet. 2011. "Is the Division of Labour Limited by the Extent of the Market? Evidence from French Cities." *Journal of Urban Economics* 69 (1): 56–71.

Duranton, Gilles, and Diego Puga. 2004. "Micro-foundations of Urban Agglomeration Economies." In *Handbook of Regional and Urban Economics*, Vol. 4, edited by J. Vernon Henderson and Jacques-François Thisse, 2063–2117. Amsterdam: Elsevier.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70 (5): 1741–79.

Ellison, Glenn, Edward L. Glaeser, and William R. Kerr. 2010. "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns." *American Economic Review* 100 (3): 1195–1213.

Fajgelbaum, Pablo D., and Edouard Schaal. 2020. "Optimal Transport Networks in Spatial Equilibrium." *Econometrica* 88 (4): 1411–52.

Fortheringham, A. S., and M. E. O'Kelly. 1989. *Spatial Interaction Models: Formulations and Applications.* Dordrecht: Kluwer Academic Publishers.

Fujita, Masahisa, and Hideaki Ogawa. 1982. "Multiple Equilibria and Structural Transition of Non-monocentric Urban Configurations." *Regional Science and Urban Economics* 12 (2): 161–96.

Glaeser, Edward, and David Cutler. 2021. *Survival of the City: Living and Thriving in an Age of Isolation.* New York: Penguin Press.

Glaeser, Edward L., Jed Kolko, and Albert Saiz. 2001. "Consumer City." *Journal of Economic Geography* 1 (1): 27–50.

Green, David R. 1988. "Distance to Work in Victorian London: A Case Study of Henry Poole, Bespoke Tailors." *Business History* 30: 179–94.

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti. 2010. "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings." *Journal of Political Economy* 118 (3): 536–98.

Gupta, Arpit, Stijn Van Nieuwerburgh, and Constantine Kontokosta. 2022. "Take the Q Train: Value Capture of Public Infrastructure Projects." *Journal of Urban Economics* 129: 103422.

Haughwout, Andrew, James Orr, and David Bedoll. 2008. "The Price of Land in the New York Metropolitan Area." *Current Issues in Economics and Finance* 14 (3): 1–7.

Heblich, Stephan, Stephen J. Redding, and Daniel M. Sturm. 2020. "The Making of the Modern Metropolis: Evidence from London." *Quarterly Journal of Economics* 135 (4): 2059–2133.

Helpman, Elhanan. 1998. "The Size of Regions." In *Topics in Public Economics: Theoretical and Applied Analysis*, edited by David Pines, Efraim Sadka, and Itzhak Zilcha, 33–54. Cambridge, UK: Cambridge University Press.

Hulten, Charles R. 1978. "Growth Accounting with Intermediate Inputs." *Review of Economic Studies* 45 (3): 511–18.

Kleinman, Benny, Ernest Liu, and Stephen J. Redding. 2023. "Dynamic Spatial General Equilibrium." *Econometrica*, forthcoming.

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.

Lucas, Robert E., Jr., and E. Rossi-Hansberg. 2002. "On the Internal Structure of Cities," *Econometrica* 70 (4): 1445–76.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences.* Cambridge, MA: Harvard University Press.

Marshall, Alfred. 1920. *Principles of Economics.* London: Macmillan.

McDonald, John F., and Daniel P. McMillen. 2010. *Urban Economics and Real Estate: Theory and Policy.* 2nd ed. Hoboken, NJ: John Wiley and Sons.

McFadden, Daniel. 1974. "The Measurement of Urban Travel Demand." *Journal of Public Economics* 3 (4): 303–28.

Michaels, Guy, Ferdinand Rauch, and Stephen J. Redding. 2018. "Task Specialization in U.S. Cities from 1880 to 2000." *Journal of the European Economic Association* 17 (3): 754–98.

Mills, Edwin S. 1967. "An Aggregative Model of Resource Allocation in a Metropolitan Area." *American Economic Review* 57 (2): 197–210.

Miyauchi, Yuhei, Kentaro Nakajima, and Stephen J. Redding. 2022. "The Economics of Spatial Mobility:

Theory and Evidence Using Smartphone Data." NBER Working Paper 28497.

Monte, Ferdinando, Stephen J. Redding, and Esteban Rossi-Hansberg. 2018. "Commuting, Migration and Local Employment Elasticities." *American Economic Review* 108 (12): 3855–90.

Muth, Richard. 1969. *Cities and Housing: The Spatial Patter of Urban Residential Land Use.* Chicago: University of Chicago Press.

Owens, Raymond, III, Esteban Rossi-Hansberg, and Pierre-Daniel Sarte. 2020. "Rethinking Detroit." *American Economic Journal: Economic Policy* 12 (2): 258–305.

Redding, Stephen J. 2016. "Goods Trade, Factor Mobility and Welfare." *Journal of International Economics* 101: 148–67.

Redding, Stephen J. 2022a. "Trade and Geography." In *Handbook of International Economics*, Vol. 5, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 147–217. Amsterdam: North-Holland.

Redding, Stephen J. 2022b. "Suburbanization in the USA, 1970–2010." *Economica*, 89 (S1): S110–36.

Redding, Stephen J. 2023. "Replication data for: Quantitative Urban Models: From Theory to Data." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E185425V1.

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

Redding, Stephen J., and Daniel M. Sturm. 2008. "The Costs of Remoteness: Evidence from German Division and Reunification." *American Economic Review* 98 (5): 1766–97.

Redding, Stephen J., and Daniel M. Sturm. 2016. "Neighborhood Effects: Evidence from the Streets of London." Unpublished.

Redding, Stephen J. and Matthew A. Turner. 2015. "Transportation Costs and the Spatial Organization of Economic Activity." In *Handbook of Regional and Urban Economics*, Vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 1339–98. Amsterdam: North-Holland.

Rosenthal, Stuart S., and William C. Strange. 2004. "Evidence on the Nature and Sources of Agglomeration Economics." In *Handbook of Regional and Urban Economics*, Vol. 4, edited by J. Vernon Henderson and Jacques-François Thisse, 2119–71. Amsterdam: Elsevier.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Raymond Owens III. 2010. "Housing Externalities." *Journal of Political Economy* 118 (3): 485–535.

Smith, Adam. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations.* London: W. Strahan and T. Cadell.

Tsivanidis, John N. 2018. "The Aggregate and Distributional Effects of Urban Transit Infrastructure: Evidence from Bogotá's TransMilenio." PhD diss. University of Chicago.

United Nations. 2019. *World Urbanization Prospects: The 2018 Revision.* New York: United Nations Department of Economic and Social Affairs.

Achieving Universal Health Insurance Coverage in the United States: Addressing Market Failures or Providing a Social Floor?

Katherine Baicker, Amitabh Chandra, and Mark Shepard

Among developed nations, the United States stands as an outlier in health insurance coverage: almost all other high-income countries have near-universal coverage, while almost 10 percent of the US population is uninsured. Figure 1 shows uninsured rates since the 1960s in the United States relative to a group of 19 developed nations. While uninsurance has declined everywhere over time, the United States has been an outlier since at least the 1980s, with most others achieving universal insurance by 1995.

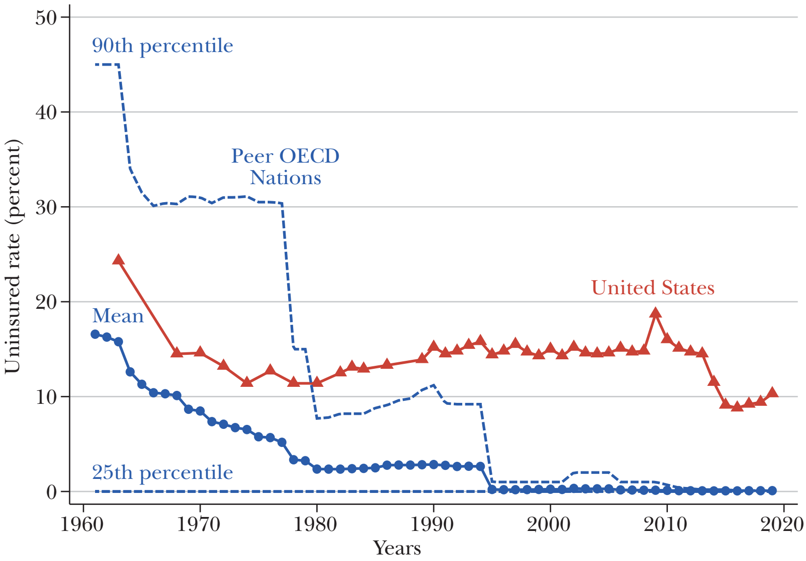
The lack of universal coverage presents a puzzle in standard economic models. Risk-averse people should benefit from some amount of health insurance, even as a purely financial product to protect against medical expense risk. Beyond financial protection, ample evidence shows that health insurance provides greater access to beneficial care and can improve health and save lives. Why, then, is uninsurance such a persistent challenge in the United States? Why is the US experience with uninsurance different from other high-income nations?

We present two approaches to understanding less-than-universal health insurance coverage in the United States and their implications for policies to expand

■ *Katherine Baicker is Dean and Emmett Dedmon Professor, University of Chicago Harris School of Public Policy, Chicago, Illinois. Amitabh Chandra is Ethel Zimmerman Wiener Professor of Public Policy, Harvard Kennedy School, and Henry and Allison McCance Professor of Business Administration, Harvard Business School, Cambridge, Massachusetts. Mark Shepard is Associate Professor, Harvard Kennedy School, Cambridge, Massachusetts. Their email addresses are kbaicker@uchicago.edu, amitabh_chandra@harvard.edu, and mark_shepard@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.37.2.99.

Figure 1.
Health Uninsurance Rate in the United States and Peer OECD Nations, 1961–2019



Source: For international data, see OECD (2022). US data is from Furman and Fiedler (2014), who compiled it from the National Health Interview Survey and other sources.
Notes The graph shows the share of populations who lack health insurance coverage in the United States and peer nations. Peer OECD nations are the 19 countries with consistent uninsurance data from 1961–2019 in the OCED Health Statistics database: Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, Germany, Iceland, Ireland, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Sweden, Switzerland, and the United Kingdom. We report the mean, 25th, and 90th percentiles across these nations.

coverage. The first—rooted in the US experience and the economics of supply and demand for health insurance—focuses on the market failures that limit availability of valuable insurance products and the behavioral frictions that further reduce take-up. This "market failures" approach has yielded a large body of fruitful research elucidating the (many) problems affecting insurance markets. However, we argue that it has been less fruitful as an effective guide to universal coverage. Fundamentally, it suggests an incremental approach to insurance expansion via targeted policies to correct market failures that inhibit take-up. The result in the United States has been a patchwork of policies, such as expanding program eligibility, increasing subsidies, streamlining or nudging enrollment, fine-tuning risk adjustment, and penalizing uninsurance through individual and employer mandates. Indeed, many of the policies in the 2010 Patient Protection and Affordable Care Act are based on this approach.

While based in part on the understandable goal of avoiding disruption of people's existing coverage, this incremental approach has sustained a fragmented

US insurance system with many inherent limitations. These include labor market "job lock" (in which workers remain in a job for fear of losing health insurance), regressive financing, costly complexity, and limited incentives for investing in population health. Even more fundamentally, this approach does not coherently define the social welfare goal of how much insurance (and health care access) should be available "universally," nor the effect of insurance design on system-level investment in medical capacity and innovation.

An alternative approach would start with universal coverage of some kind as a social goal and focus on the decisions involved in designing a health insurance system that ensures a floor on access to medical care. This approach—related to the path taken by many other high-income countries—automatically provides a basic level of insurance to everyone and then focuses attention on key questions about the design of basic coverage and the availability of alternatives.

This "social floor" approach makes explicit many of the underlying goals and tradeoffs that are obscured in the incremental approach grounded in correcting market failures. While take-up of (basic) insurance is no longer a core issue—because everyone gets basic coverage automatically—economics can play a key role in framing the problems and understanding tradeoffs that arise.

We highlight three key questions that arise in the social floor approach. First, this approach to universal coverage requires defining the floor to which everyone will be automatically entitled, or what we call the "basic bundle." Defining this scope of coverage requires a difficult public conversation: not an abstract debate about whether "health care is a right," but an answer to the concrete question of "*how much* health care is a right" given real-world funding, capacity, and resource constraints. This process starts by defining what set of medical services are covered, but it must go further. Almost all health services can be "medically necessary" for certain patients in certain situations but quite wasteful (with virtually no health benefit) in other situations. The generosity of basic coverage depends on which mechanisms are used to limit spending on covered services—such as global medical budgets, provider prices, capacity constraints, patient cost sharing, and utilization controls.

Second, a social floor approach must specify *who is in charge* of administering and delivering basic coverage—who decides how much to pay for which services for which patients? In the current US system, some of these decisions are delegated to private insurers, while others are subject to federal and state regulation, leading to different degrees of choice, flexibility, and alignment with patient preferences across insurance segments. A more coherent system for administering basic coverage could yield benefits of simplicity and lower administrative costs.

Finally, decisions must be made about whether and how individuals can use private funds to obtain additional coverage beyond the basic package. This decision about allowing "top up" has economic as well as ethical and distributional implications. The more heterogeneous the population in terms of preferences and income, the greater the return to additional choices, but also the wider the resulting disparities in outcomes.[1] In addition, a top-up system will increase incentives for innovators to invest in new treatments, given a large monopsony purchaser catering to median preferences. (The ability of the US government to demand monopsony prices will likely exceed that of other smaller governments, and the evidence for sustaining a "moderate monopsonist" is weak.)

Beyond these first-order questions, myriad political and logistical concerns would arise in moving the United States to a different framework—though there are ways to smooth that transition path so that changes are not unduly disruptive, as we sketch out below. There are lessons to be learned from the experiences of other countries, many of which have some flavor of a universal basic system, though with different answers to the fundamental questions posed above. For example, the United Kingdom automatically covers all residents in its National Health Service, a public healthcare system with no out-of-pocket costs. The Netherlands and Switzerland provide universal coverage through a health insurance market in which people can choose among competing private plans offering basic coverage. Germany and Israel have systems of basic coverage through competing nonprofit plans. Australia has a basic public medical system like the UK system, but with a much larger role for private hospitals and insurance. In many countries, employers play a central role in providing top-up coverage. While these designs display considerable variation, they share a common feature that all citizens are automatically entitled to a basic level of health insurance, without the need to purchase a product or go through a complex enrollment process, resulting in essentially zero uninsurance.

We begin with a short synopsis of the rationale behind a goal of universal coverage, the evolution of health insurance coverage in the United States, and comparisons to other systems. We then draw out implications for coverage through an approach grounded in addressing market failures within the current system versus starting with a foundation of a guaranteed coverage floor.

Evolution of Health Insurance Coverage

Rationales and Goals of Universal Health Insurance Coverage

We begin with a presumption that almost all individuals are risk averse, and thus inherently value at least some basic amount of insurance. Insurance coverage improves health outcomes and provides financial protection to providers and payers, as well as to covered individuals: for reviews of the evidence, see Finkelstein, Mahoney, and Notowidigdo (2018) and Sommers, Gawande, and Baicker (2017). Moreover, the value of health insurance increases as medical technology advances and more lifesaving but expensive treatments become available—for example, a gene-therapy that allows children crippled by spinal muscular atrophy

[1] Americans are quite divided in how they approach basic questions about government intervention in health care and the degree to which they value others' insurance coverage (Baicker and Chandra 2020).

to walk, or a cell therapy that edits DNA to neutralize genes that cause heart attacks.

We also assume that society places value on providing health insurance to others, which could arise for several reasons.[2] One societal motivation for expanding coverage to the uninsured is risk of health spillovers; but although the COVID pandemic represents a recent example of enormous health spillovers, such spillovers are thought to be relatively small in normal times. Another motivation stems from altruistic concerns for the health of others, especially if health shocks are seen as largely exogenous, or if patients are seen as underconsuming health care because of "behavioral hazard" (Baicker, Mullainathan, and Schwartzstein 2015). Finally, conditional on a social decision to provide life-saving care regardless of ability to pay, there is a social interest in providing that care efficiently: the uninsured impose costs on others when they use inefficient "uncompensated care" in emergency rooms and safety net hospitals (the Samaritan's Dilemma). Further, these costs increase with growth in lifesaving medical technology that is expensive.[3]

Establishing the private and social rationale for all residents to have *some* health insurance leads next to the question of *how much* health insurance. A truly unlimited right to health care (that is, any care at any price for anyone) would quickly eat up all resources available for all other public programs, including schools, housing, and public health. Before turning below to different mechanisms for establishing limits to coverage and spending, we first address the question of why there remains such a substantial population with no insurance at all in the current US system.

The Development of the US Health Insurance System

Prior to the twentieth century, few people in any country had formal health insurance. Medical care was not effective or expensive enough to motivate an insurance product to cover its costs. As medicine advanced and became more expensive, the value of health insurance grew. Starting with Otto von Bismarck's Germany in 1883, many high-income countries developed social health insurance systems that covered wide swaths of the population, often through employers or workers guilds. From the 1940s to 1970s, these social insurance systems gradually expanded or evolved into national health insurance systems in many countries. As Figure 1 indicates, most US peer high-income nations had near-universal coverage by 1980. By 1995, universal coverage had come to nearly all peer nations.

[2] Some would argue that the choice to remain uninsured is a matter of individual liberty; that is, the freedom not to purchase a product. This argument has been encapsulated through comparing mandating universal insurance coverage to mandating that people eat broccoli (for example, Elhauge 2011). Like broccoli, insurance is good for health—but should the government therefore mandate it? This argument is strongest if insurance coverage is viewed as a purely private good.

[3] Finkelstein, Hendren, and Luttmer (2019) estimate using the Oregon Health Insurance Experiment that third-party uncompensated care costs equal about two-thirds of the cost of formal insurance via Medicaid. Mahoney (2015) estimates that the Pigouvian externality of unpaid medical debts discharged in bankruptcy (just one part of uncompensated care) are about $340 per person.

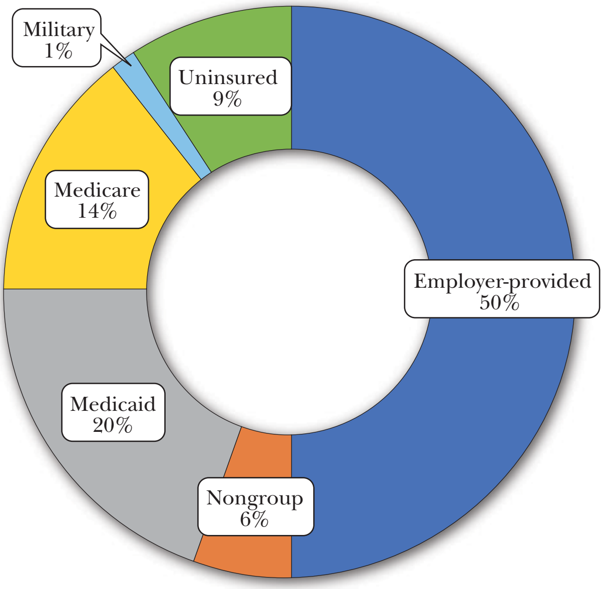
Unlike other high-income nations, the United States did not implement a single model of public (or private) health insurance. Instead, it developed a patchwork of programs for different groups. The United States gradually adopted an employer-based health insurance system over the first half of the twentieth century. Such plans received a major boost from the decision—made in the throes of World War II—that workers could receive raises (during a time of wartime wage controls) in the form of employer-paid health insurance excluded from personal taxable income. Additionally, families could choose to purchase private "nongroup" coverage directly from insurers. However, by the early 1960s, about 25 percent of Americans lacked health insurance, and these were disproportionately elderly retirees and low-income people.

The first major public insurance expansion occurred with the creation of Medicare and Medicaid in 1965. These programs were designed to cover the key groups with the highest uninsured rate: the elderly and families with low incomes, as well as people with disabilities (added to Medicare in 1972). As a result, the national uninsured rate fell from 25 percent in 1963 to 11–12 percent by the mid-1970s.

But over the four decades from 1973 to 2013, the United States made little net progress in reducing the uninsured rate. The uninsured rate ticked up to 15 percent during the 1980s and remained around or above that level until 2013. This standstill occurred despite the growth of Medicaid to cover more low-income pregnant women, parents, and children—especially after the passage of the Children's Health Insurance Plan (CHIP) in 1997—gains that were roughly offset by declines in employer-provided insurance. It also occurred despite a major expansion in public spending on Medicare and Medicaid, which rose from $13 billion (or 17 percent of national health expenditures) in 1973 to $1 trillion (or 37 percent of national health expenditures) by 2013 (US Centers for Medicare and Medicaid Services 2022). The value of the tax exclusion for employer-sponsored health insurance also grew to $270 billion in foregone income and payroll taxes (Tax Policy Center 2022). Yet despite these major expansions in public spending and eligibility, uninsurance did not fall meaningfully.

The Patient Protection and Affordable Care Act (ACA) that became law in 2010 represented the second major wave of coverage expansion within the current system. The law provided a plausible path to universal coverage, at least for citizens. It expanded Medicaid to everyone with incomes below 138 percent of the federal poverty level (in states that adopted the Medicaid expansion) and provided income-based subsidies for private insurance in newly created health insurance exchanges. Nearly all poor and middle-income citizens—those with incomes below 400 percent of the poverty line, or $92,000 for a family of three—qualified for either Medicaid or subsidized insurance at premiums at a cost of 2–10 percent of income for a benchmark plan. Higher-income Americans did not generally qualify for subsidies, but they were given access to a minimum standard of insurance on newly created state exchanges and encouraged to take it up through a tax penalty on uninsurance (though this was repealed in 2019). When the ACA insurance expansions took effect in 2014, uninsurance rates dropped from about 15 percent down to 9–10 percent—about 30 million uninsured people.

Figure 2
US Health Insurance Coverage by Source, 2019



Source: Kaiser Family Foundation (2021b) analysis of American Community Survey data.
Note: The figure shows the shares of the overall US population with health insurance from each source of coverage.

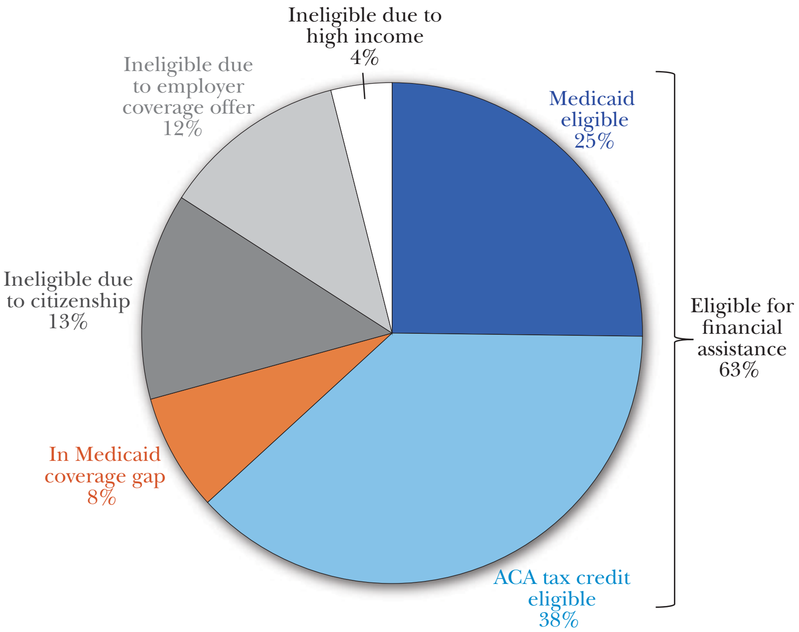
Figure 2 shows population shares in various forms of health insurance as of 2019, based on data from the American Community Survey. Half of the US population (158 million people) had employer-provided insurance, while one-third had either Medicare (45 million) or Medicaid (63 million), though the latter has risen sharply since the start of the pandemic. Another 6 percent (19 million) had nongroup coverage (including coverage in the health insurance exchanges created in the 2010 legislation), up slightly from the 5 percent share prior to 2014. Finally, 9 percent (29 million) lacked formal health insurance.

Explaining the Persistence of Uninsurance

What explains stubbornly persistent uninsurance in the United States? Much of the public discourse focuses on affordability. However, although available insurance may be too expensive for some to buy, an examination of the data suggests this is unlikely to be the whole story.

Figure 3 breaks down the uninsured into shares eligible for various sources of insurance as of 2021. On the one hand, this figure points to some gaps in social safety net programs. Not all states have expanded Medicaid under the Patient Protection and Affordable Care Act of 2010, leaving about two million very low-income

Figure 3
Eligibility for Subsidized Insurance Coverage among Nonelderly Uninsured, 2021



Source: Kaiser Family Foundation (2021a), using estimates from American Community Survey data. The graph follows the format of KFF in Tolbert, Orgera, and Damico (2019).
Note: The graph shows the share of uninsured Americans under age 65 who are already eligible for subsidized insurance via Medicaid or ACA tax credits to purchase coverage on state insurance exchanges. Overall, 63 percent of the uninsured are eligible for financial assistance, while 37 percent are not. The "Medicaid coverage gap" refers to low-income individuals living in states that have not expanded Medicaid under the Patient Protection and Affordable Care Act of 2010 (ACA).

Americans in twelve states to fall into a "coverage gap" (not eligible for Medicaid, but too low income to be eligible for nongroup market subsidies). Further, undocumented immigrants are not eligible for subsidies or Medicaid under the 2010 law, affecting perhaps 4 million people (or 13 percent of the uninsured). But together, these two groups account for less than one-fourth of the remaining uninsured.

About 63 percent of the uninsured (about 18 million people)—by far the largest share—are low- or middle-income Americans who qualify for subsidized insurance (via Medicaid or a health insurance exchange) that they have not taken up. Indeed, under the more generous subsidies available since 2021, about 40–50 percent of the uninsured likely qualify for fully-subsidized coverage; that is, coverage with zero out-of-pocket premium for them (Rae et al. 2021). Thus, a substantial share of the uninsured could be covered by inducing take-up of benefits that would be *free* to them. These facts indicate that affordability is not the only, or even the main, barrier to universal coverage; other forces are at work as well.

Approaching Universal Coverage by Addressing Insurance Market Failures

If the reason uninsurance persists is not merely unaffordability for credit-constrained low-income populations, the standard model suggests an examination of potential insurance market failures. What are the right policy responses to correct market failures and align incentives for take-up?

This "market failures approach" to universal coverage is the (largely implicit) workhorse in much of the relevant economics literature. This framework starts by conceptualizing health insurance as a product bought by consumers who obtain value from it (*demand*), and sold by insurers who incur costs in selling the policy and covering the care (*supply*). By standard theory, the forces of supply and demand should lead to Pareto optimal allocations unless there are market failures or behavioral frictions. This territory is familiar and comfortable for economists.

The logic of risk aversion and uncertain health expenses suggests that most (perhaps all) consumers should benefit from purchasing a nonzero amount of insurance.[4] Therefore, if many consumers lack *any* (formal) insurance, it is natural to ask whether the outcome is Pareto optimal—and if not, what the problems are and how to fix them. Over the past decades, economists have elucidated a long list of factors that may lead to nonoptimal uninsurance. Here, we review them briefly, grouped into four categories.

First, health insurance markets suffer from *adverse selection*. In addition to truly asymmetric information about consumer health risk, existing regulations ban health insurers from price discriminating based on much of the information they do have about individual-specific risks. Instead, insurers must use group average costs to set premiums. As a result, low-cost healthy individuals are charged premiums exceeding their own expected costs—because these include a cross-subsidy for sicker individuals—and may find purchasing health insurance to be a bad deal. The implications of adverse selection are carefully drawn out in the theory literature, and the past two decades have seen a burgeoning of empirical work showing its continued relevance; for some useful starting points to the modern literature, see Einav and Finkelstein (2011) and Geruso and Layton (2017), both in this journal. Recent work, however, suggests that adverse selection may not be sufficient to explain low take-up (at least among the poor), because a large share of low-income individuals have demand for insurance falling far below their costs of coverage (Finkelstein, Hendren, and Shepard 2019).

Second, the *presence of insurer market power and/or loading fees* to cover administrative expenses may discourage individuals from purchasing insurance. These forces push premiums above actuarially fair levels, meaning that consumers again may find it to be a bad deal. Insurance markets are highly concentrated, and a growing body of work shows the relevance of insurers' market power on premiums (for

[4] This positive insurance result persists even with moral hazard, as long as the cost of moral hazard for the first unit of insurance is second-order, while the benefits of risk protection are first order.

example, Dafny 2010; Dafny, Duggan, Ramanarayanan 2012; Starc 2014; Mahoney and Weyl 2017).

Third*, behavioral frictions combined with liquidity constraints* may discourage consumers from obtaining health insurance, because such consumers depart from the rational agents with easy access to capital markets who are the starting point of standard economic theory. Relevant factors include liquidity constraints (Ericson and Sydnor 2018); biased beliefs about health risks (Spinnewijn 2017); information frictions (Domurat, Menashe, and Yin 2021); and inertia in the face of enrollment hassles (Shepard and Wagner 2022). A growing body of evidence finds that even when consumers *do* purchase insurance, they often choose poorly (Abaluck and Gruber 2011; Handel 2013; Bharghava, Loewenstein, and Sydnor 2017). Similarly, as patients they often make imperfect medical decisions in the face of cost sharing (Newhouse 1993; Baicker, Mullainathan, and Schwartzstein 2015; Brot-Goldberg et al. 2017).

Finally, the *presence of an implicit safety net* providing health care for the uninsured may undermine the incentives of some individuals to pay for insurance (as in the "Samaritan's dilemma" discussed by Buchanan 1975). We have, as a society, already made the decision that vital care must be provided to people in critical need of care, regardless of ability to pay. Informal safety net coverage goes beyond requirements that emergency departments address critical needs regardless of ability to pay, like those embodied in the Emergency Medical Treatment and Active Labor Act of 1986. There is also "charity care" delivered by a range of providers and informal insurance from family and friends (Finkelstein, Mahoney, and Notowidigdo 2018). Because of the relatively low threshold for bankruptcy (Mahoney 2015) and free or discounted care from safety net providers (Garthwaite, Gross, and Notowidigdo 2018), third parties cover about 80 percent of the costs of many low-income uninsured (Finkelstein, Hendren, and Luttmer 2019). Thus, even uninsured Americans have a sort of informal health insurance coverage—albeit coverage that is disorganized, stressful, low quality, and inefficient.

Addressing Market Failures to Expand Health Insurance Coverage

With an approach rooted in market failures, the natural response is to implement targeted policies that address those failures. We describe how four incremental policy approaches might work within the system: expanding eligibility, expanding subsidies, encouraging enrollment in health insurance for those who already qualify, and bolstering safety net care.

One of the most straightforward expansions of health insurance eligibility within the current system is *to expand Medicaid eligibility in states that have not done so*, using the heavy federal subsidies included for this purpose in the Affordable Care Act. This would expand eligibility for health insurance to about two million people.[5] A reason commonly stated by states that have not taken this step is a concern that

[5] For estimates of the health uninsured cited in this section, see estimates from the Kaiser Family Foundation (2021a).

the federal subsidies will be withdrawn in the future, which would lead state-level politicians to face an unpalatable choice between finding an alternative funding source or cutting benefits. In our view, however, the choice not to expand seems more a matter of politics than of economic calculus. Additional expansions could also cover the four million uninsured who are ineligible because of immigration status, though such a step is even more politically fraught. (We consider "Medicare for All" proposals to be a more fundamental system change, addressed below.)

Increasing the generosity of subsidies for those purchasing nongroup private insurance may increase enrollment, especially among healthy near-poor (150–300 percent of poverty) individuals for whom existing modest premiums (about 2–6 percent of income) may nonetheless impose a significant barrier to take-up (Finkelstein, Hendren and Shepard 2019; Tebaldi 2022). Some groups are already eligible for partial subsidies but may find (or at least perceive) insurance to remain unaffordable.

Addressing the frictions associated with enrolling in and retaining health insurance may substantially increase coverage by inducing those who already qualify for health insurance coverage at no out-of-pocket cost or with heavy subsidies to take it up. This group comprises about 22 million of the 29 million uninsured, including about 7.3 million people who already qualify for free Medicaid, 11.0 million people who qualify for health insurance through a state-level insurance "exchange" (with about half that group qualifying for free coverage), and 3.5 million people who could be receiving health insurance through an employer. For example, recent work has highlighted how seemingly small administrative burdens involved with insurance enrollment can strongly affect coverage outcomes (Domurat, Menashe, and Yin 2021; McIntyre, Shepard, Wagner 2021; Shepard and Wagner 2022; Wright et al. 2017). This is especially true when individuals transition between different forms of coverage—for instance, between Medicaid and exchange eligibility, or after losing a job with employer coverage and qualifying for Medicaid. Policies that target transitioning individuals for outreach or auto-enrollment could have a significant impact on take-up, though there are implementation challenges (Dorn, Capretta, and Chen 2018).

Finally, *the existing safety net system of emergency departments, federally qualified health centers, public clinics, and charity care could be bolstered.* Eligibility along with the bundle of free services expected to be delivered could be expanded. For example, providers of such services could be reimbursed with public funds for primary care and medicines that prevent acute events, not just for emergency department visits.

Limitations to the Approach of Expansion through Filling Gaps in Current System

The US health insurance system has a number of well-documented issues beyond the gaps in health insurance coverage. One is the frictions in labor markets introduced by the fact that half of the population (157 million) is covered by employer-based health insurance. People know that if they lose their jobs—because of recessions, pandemics, or business failure—they lose their health insurance, which likely also means finding a new primary care physician, transferring medical records, and amending medications to conform with formularies. This generates "job lock" that reduces labor market flexibility (Madrian 1994).[6]

Another limitation is the lack of continuity of care and coverage introduced by the discontinuities in eligibility between different forms of health insurance—including employer-based insurance, Medicaid, and the subsidized health insurance exchanges for individual policies. In general, a multipayer system is also expensive to administer: each payer has its own reimbursement forms that are not standardized, and there are bespoke cost sharing, networks, and formularies—which imposes costs and can confuse patients and their doctors.

Approaches grounded in addressing market failures in the current system are perhaps the path of least resistance in the short run, minimizing disruptions to care while marginally increasing coverage. It is worth noting, however, both the limited effectiveness of such approaches over the last 50 years and the shortcomings that such patches would perpetuate.

Universal Coverage through Establishing a Social Floor

Rather than beginning with the presumption that the main need is addressing market failures, an alternative approach to expanding coverage begins with the explicit presumption that covering everyone with some form of insurance is a social goal.

In every nation, citizens have some access to health care, regardless of ability to pay, simply by being part of society—the "right" to a de facto floor of care. The United States also has an implicit floor, albeit an informal one, meaning that even the uninsured have access to some health insurance, with no credible way to opt out. We call this the "basic bundle." The US basic bundle includes hospital care in emergencies as required by the Emergency Medical Treatment and Labor Act of 1986, and nonemergency care from community health centers, safety net hospitals, and clinics that treat people regardless of insurance status or ability to pay. It is socially costly, involving about $40 billion in annual uncompensated care and $11 billion in grants for community health centers, paid for by a mix of public funding and health system cross-subsidies.[7] In this way, the US basic bundle of health care is not unlike public health systems available in many developing countries that are principally used by the poor.

[6] It is worth noting that the tax financing of employer-based insurance is inherently regressive and inefficient. By making employer-sponsored health insurance policies tax-exempt, the largest benefits go to workers in the highest tax brackets with the most generous policies. Such policies may also foster low-cost-sharing and higher-premium plans, exacerbating moral hazard issues.

[7] For data from the Kaiser Foundation on "Sources of Payment for Uncompensated Care for the Uninsured," see https://www.kff.org/uninsured/issue-brief/sources-of-payment-for-uncompensated-care-for-the-uninsured/. For data on Community Health Center Revenues by Payer Source, see https://www.kff.org/other/state-indicator/community-health-center-revenues-by-payer-source.

The implicit basic bundle could be made explicit through automatic, free enrollment in some form of coverage financed by general revenues. This broad approach is taken in nearly all countries that have achieved universal coverage, but there are many variations. Although "universal coverage" is often equated with single-payer, government-run health insurance, systems in peer nations in fact reflect a diversity of models with varying roles for government.[8] For instance, the Netherlands and Switzerland provide insurance via *universal health insurance markets*, offered by competing (but regulated) private health insurers. Germany provides coverage via *competing nonprofit insurers* called "sickness funds" that offer standardized benefits and cover the same set of providers (with common fee schedules). Germans can also opt out into a less regulated private health insurance market, an option taken by 11 percent of (mostly higher-income) people.

Canada and the United Kingdom both have universal coverage through *single-payer government-run health insurance*. However, the United Kingdom's medical provider system is also government-run, whereas Canada's providers are largely private. Further, despite being "single-payer" systems, both nations feature a sizable role for add-on private insurance (largely provided through employers) to cover extra services. In the United Kingdom, 11 percent of people have private insurance that covers supplementary benefits—largely elective care at private hospitals with shorter waits. In Canada, 67 percent of people hold complementary private insurance that covers services excluded from the public plan (for example, prescription drugs and dental care).

Many components of the United States's current patchwork system have parallels to international health insurance models; for example, traditional fee-for-service Medicare is similar to Canada, the state-level health insurance exchanges are similar to Switzerland, and the Veteran's Health Administration is analogous to the United Kingdom's system. Thus, moving towards one of these models need not involve wholesale overhaul. But there must be explicit policy decisions made on multiple dimensions that are only implicitly determined now.

We discuss three key policy decisions in a system of guaranteed universal basic coverage: (1) What health care does the basic bundle cover, and how generous is that coverage? (2) What mechanisms are used to limit spending, and who decides? (3) Are people permitted to purchase top-up or supplementary coverage beyond the basic bundle? One goal of this article is to provide a framework that may guide future research to help inform answers to these questions.

Design Question #1: What Does the Basic System Cover?

How generous—and therefore expensive—should the basic bundle be? This question has important implications for the level of health spending and the ultimate disparities in health care and outcomes, and the answer is a matter of public

[8] For an overview and sources of information on these systems, see the Commonwealth Fund International Health System Profiles (2022) at https://www.commonwealthfund.org/international-health-policy-center/system-profiles.

policy priorities and preferences. We argue that the key input into this social welfare function should be the value of the care in improving health relative to the resource cost of the care. Care with health benefits that sufficiently exceed resource costs should be included.[9]

It is important to note that "high-value care" does not mean "low-cost care": some very expensive treatments with dramatic health benefits are high-value, and some cheap treatments with negligible health benefits are low-value. Some health care services are of such high value that they have negative net cost—that is, the service *pays for itself.* This small minority of care could include vaccinations against communicable diseases, superior treatments for mental illness that reduce incarceration of patients with schizophrenia, or future novel transformational treatments for diseases like Alzheimer's that reduce total spending. Some health care is of so low value that it has negative net benefit—that is, it is *harmful* to patients. This too is only a small share of care, like prescribing antibiotics for viral infections or contraindicated MRI scans.

But most health care has a positive cost that must be weighed against a positive health benefit. Lots of care has health benefit that will clearly warrant its cost to most: say, emergency care for acute events like accidents, strokes, appendicitis, or pulmonary embolisms, or "curative" or life-sustaining medicines. Coverage of such treatments in the "basic bundle" would likely be uncontroversial. But this leaves a host of care with high cost and more questionable benefits, and debate about inclusion of such services in the basic bundle would likely be heated. As discussed below, establishing a regularized mechanism for inclusion decisions about whether care has sufficiently high benefits relative to costs is important for a successful policy—and something that many countries have struggled to achieve.

Such a system would not only focus health care resources on high-value care, but would also provide an incentive for innovators to develop new treatments with higher health benefit and/or lower cost. The ideal health insurance system should not only provide efficient coverage for today's technology, but should also embed appropriate incentives for the development of meaningful innovations in future medical care including prevention, delivery, devices, medicines, and procedures.

In turn, insurance coverage must evolve in response to innovation in care. For example, Medicare only began covering prescription drugs in 2006—a relic of the fact that such medications had not been an important or expensive component of care when the program was established in 1965.[10] The design of public health

[9] Additional criteria for inclusion in the basic bundle might include services for which the top-up markets discussed below are unlikely to function well because of adverse selection, or services that are disproportionately used by disadvantaged populations where there is high distributional social value in coverage.

[10] Canada has "universal coverage" in the sense that everyone has coverage for hospital and physician services, but 20 percent of Canadians lack prescription drug coverage. In the United States, standard Medicare does not cover vision and dental benefits, and the Medicare drug coverage long had an infamous "donut hole" where patients lose insurance protection—a design artifact that is believed to have increased mortality as patients cut back on their medicines in response to this gap in coverage (Chandra, Flack, and Obermeyer 2021). Private health insurance plans offered drug coverage three decades before Medicare.

insurance—from coverage to reimbursement rates to gatekeeping mechanisms—is a major driver of investment in capacity as well as innovation (for example, Finkelstein 2007; Clemens and Gottlieb 2014; Clemens, Gottlieb, and Hicks 2021; Weisbrod 1991; Chandra and Skinner 2012). Incentives for innovation are not usually contemplated in "public utility" approaches to insurance regulation, and public plans tend to lag private insurance in coverage of health care innovations.[11] These facts speak to the value of allowing "top-up" plans, described below, as well as the importance of a mechanism for ensuring regular updates to basic coverage design.

Design Question #2: What Mechanisms Are Used to Control Spending?

In addition to the fundamental generosity of coverage and design of the basic bundle, decisions must be made around *mechanisms to control spending* such as cost-sharing rules, provider payments, access to provider networks, and utilization controls like prior authorization and step therapy (that is, trying less expensive options before stepping up to more expensive ones). These are often detailed decisions that cannot be specified in law but need to be made for thousands of specific instances. A key governance question naturally arises: *who is in charge* of making these detailed choices, and through what process?

It is tempting to side-step the issue of the need to control spending by suggesting that we can fund universal coverage by eliminating waste, fraud, and abuse; or by eliminating private sector profits; or by reducing administrative costs.[12] Aside from the limited magnitude of such potential savings, such arguments miss the inherent opportunity cost of spending on care with diminishing returns. Everyone is against fraud, but even assuming that we could identify in advance and prevent all "wasted" care, that would still leave an enormous body of care with limited health benefit and high cost that would eat up an increasing share of GDP as medical innovations arrived. With scarce resources, there is an inherent tradeoff between covering

Medicare's coverage of prescription drugs for over 45 million elderly Americans increased innovation in medicines that disproportionately helped these covered patients (Blume-Kohout and Sood 2013).

[11] There are global general equilibrium effects to the decision made by a large, high-income country like the United States: coverage and pricing decisions in US markets drive the development of innovations that are then available to other countries—in essence cross-subsidizing innovations that benefit citizens of other countries, but also driving potential expenses for their systems.

[12] Versions of this argument include asserting that Medicare has low administrative costs relative to private insurers. Both medical prices and administrative costs ("paperwork") appear to be much higher in the United States than in comparable countries. For medical prices, see Anderson, Hussey, and Petrosyan (2019). For administrative costs, see Cutler and Ly (2011). We say "appear to be" because cross-country price comparisons are notoriously challenging, as it is difficult to define a constant, quality-adjusted unit of service. However, not all administrative costs are wasteful, as many involve efforts to limit use of high-cost drugs and treatments (Brot-Goldberg et al. 2022). Administrative costs are low in Medicare partly because it does not perform utilization management in its fee-for-service offerings, relies on regulated prices as a way to manage utilization, and piggybacks off the systems used by private insurers to process claims. Although there is surely plenty of room for efficiency gains, there are also likely to be real tradeoffs.

more people and covering more resource-intensive services (Baicker and Chandra 2010).

One approach to making these choices is through a centralized public process. One possibility would be legislation that set broad guidelines to define a basic bundle and empowered a medical board or government agency to define details. It is worth noting that centralized decision making can occur within a national health insurance system largely operated by private actors. For instance, German (private) health insurance "sickness funds" have standardized coverage and cost sharing rules, and they offer essentially unrestricted choice among providers (who are paid via a centrally set fee schedule). In market-based systems like Switzerland and the Netherlands, coverage by the basic plan is universal, but people may choose different specific plans.

The benefit of this centralized approach is its simplicity and lower administrative costs. The downside is that public entities may make suboptimal decisions. On the one hand, there may be public pressures that generate unsustainably high spending (as seen in some aspects of Medicare and Medicaid), or, on the other, there may be budget pressures that generate stinting (a common perception of the United Kingdom's National Health Service, limited drug coverage in Canada, or Medicaid provider payments in many states). This fundamental problem is hard to avoid in the absence of competitive forces and market price signals.

Several tools might help to reduce the risks involved with centralized pricing or rate setting. One approach is capitation, which refers to health insurance making payments to health care providers on a (risk-adjusted) per-enrollee basis, not on a fee-for-service basis. The hope is that capitation payment provides an incentive for health care providers to innovate in ways that will attract enrollees while still holding down costs.

For example, under the Medicare Advantage program (Part C of Medicare), the government makes a flat per-enrollee payment to a private-sector insurer. There is evidence that such mechanisms drive payers to compete on quality (which direct government provision does not) and to deploy a variety of contracting arrangements with doctors and staff to reduce overuse and therefore costs (Newhouse and McGuire 2014; Curto et al. 2019). There are certainly challenges to figuring out how to risk-adjust payments to health plans so that they are incentivized neither to avoid sicker patients (Brown et al. 2014) nor to "upcode" medical diagnoses to increase payments (Geruso and Layton 2020). As another example, "Accountable Care Organizations" are groups of health care providers who provide fee-for-service care to Medicare patients, but who are financially rewarded for meeting certain predefined metrics of quality while spending less. Evidence suggests that the existing Accountable Care Organizations have generated modest cost-savings (McWilliams et al. 2018), though it is not clear how they negotiate prices for care delivered outside their own organizations.

As an alternative to centralized decision-making, there are hybrid options that vest more decision-making in private insurers, such as subjecting insurers to minimum adequacy regulations, but then giving broad flexibility to make coverage decisions, design cost-sharing schedules, and adjust provider (or pharmacy) networks. Patient cost-sharing can be a valuable tool, but it is crucial that cost-sharing take into account patients' behavioral response and align with the health value of the care to ensure that patients do not cut back on highly valuable care in response to copays (Chandra, Gottlieb, and Hicks 2021; Brot-Goldberg et al. 2017; Baicker, Mullainathan, and Schwartzstein 2015; Chandra, Gruber, and McKnight 2010). Private insurance markets can suffer from severe market failures (notably adverse selection) and suboptimal consumer choices (for example, consumers choosing low-premium plans that expose them to high patient cost-sharing), highlighting the value of policy guardrails.[13]

This hybrid approach is taken by the Medicare Part D drug benefit and by the health insurance exchanges established by the Patient Protection and Affordable Care Act of 2010. Empirically, however, it is not clear how efficient the drug coverage decisions of Part D plans are—for example, many have cost-sharing on drugs with little scope for overuse, thus reducing the insurance value without improving efficiency of resource use.

Design Question #3: What Supplementary Coverage Should Be Available?

Once the parameters of a basic, guaranteed plan are established, a policy decision needs to be made about the allowability of supplemental plans for private purchase. Supplemental plans offer several advantages—though there are important distributional implications.

First, many individuals (especially those with higher income) may wish to purchase additional coverage or access to care, and there is social value in letting people chose a plan that fits their preferences. Second, allowing for top-up insurance relieves the budgetary pressure of providing a substantially larger bundle of health care for everyone. Third, the presence of a private health insurance market can help in the process of price revelation and guide administration of the basic bundle in that way. In particular, without private markets, the regulator has no external benchmark of value, and monopsony pricing by a centralized authority risks reducing welfare by discouraging quality or investments in innovation (Chandra and Garthwaite 2019).[14] The risk of monopsony pricing increases as the share of people covered by the basic-bundle increases. Fourth, the supplemental health insurance

[13] A variation would be to include a public option to increase competition for private plans. However, private plans cannot compete with a public plan that is allowed to run massive deficits, which highlights the problems of running a system without budget limits. Traditional Medicare, for example, competes against Medicare Advantage private plans, but its deficit financing creates an unlevel playing field, limiting the market discipline that is exerted (Chandra and Garthwaite 2019).

[14] This concern is not theoretical: the prices paid to medical providers affect which providers are willing to accept patients covered that plan. Medicaid provides relatively low provider payment rates, but most Medicaid plans have limited provider networks and about 30 percent of physicians do not accept any new Medicaid patients (Holgash and Heberlein 2019). More generally, a large monopsonist payer may prefer to use administratively-set prices to control spending, which can result in a low-quality insurance product. Allowing citizens to top up the basic plan provides a signal of price adequacy—as the share of citizens with the top-up plan increases, the more likely it is that the basic plan is inadequate.

Table 1
Dimensions of Top-Up Benefits in Health Insurance

Top-Up Benefit	*Description*	*Examples*
Patient cost sharing	Health insurance systems often include cost sharing to reduce moral hazard. Individuals can purchase top-up coverage to help insure these costs.	"Medigap" insurance in the US Medicare program. Tier choice in US health insurance exchanges (platinum/gold/silver/bronze). Choice among plans with varying cost sharing in the Swiss and Dutch systems.
Add-on services	Most national health systems purposefully exclude certain categories of medical services. Common exclusions are long-term, dental, and vision care.	Outpatient prescription drugs in Canada Dental and vision coverage in US Medicare. Long-term care in many countries including the US, Canada, and the UK (except for the impoverished).
Private providers	Many health insurance systems do not cover certain providers, who may differ from others in terms of quality, convenience, or amenities.	The UK typically does not cover care at private hospitals, which offer elective procedures with shorter wait times and more amenities. About 10% of people hold private coverage to help pay for these. Many US providers do not take Medicaid and are accessible only if people pay out of pocket or purchase other insurance.
Medical amenities	The basic system often does not cover services deemed "amenities" rather than "medical quality."	In many countries, shorter waiting times for nonurgent procedures (like joint replacement surgery) are treated as an amenity. In Singapore, basic public insurance pays for shared hospital rooms or wards, while individuals who pay out of pocket or with private insurance can get private rooms in the same facility.

market can be an area for experimentation in how health care benefits might be designed or adjusted. Fifth, allowing top up of the basic bundle would not require the elimination of the existing employer-provided health insurance plans that cover about 160 million Americans. Of course, the specific choices involved in defining a basic bundle and the allowable types of top-up coverage will pose a version of the classic efficiency-equity tradeoff (Shepard, Baicker, and Skinner 2020).

A basic bundle of health care benefits could be "topped up" along multiple dimensions. Table 1 summarizes four of them (available in many national health systems): patient cost sharing, add-on services, breadth of provider network, and medical amenities. These categories highlight the dimensions along which a basic bundle would need to be defined.

There are two different potential mechanisms for top-up coverage: the ability to buy "add-on" coverage that wraps around the basic bundle, or the ability to purchase a "replacement" plan that supplants the basic program. Most high-income countries allow add-on (or "complementary") private coverage to their nationally guaranteed plan, which typically covers amenities or providers not covered in the public system (like private hospitals or private rooms within hospitals). As one example, all UK residents can use National Health Service doctors and hospitals for free, but about 11 percent of UK residents purchase private insurance that covers care at private hospitals that have private rooms, as well as shorter waiting lists for nonemergency procedures like joint replacement surgery. In other cases, patients pay out-of-pocket for higher quality treatments or amenities.

In countries that allow replacement health insurance, individuals can opt out of the baseline public insurance system to purchase less-standardized private insurance, which often features more generous treatment coverage or provider access. Prominent examples include Germany and Chile. In Germany, the tax-financed social health insurance program is the universal default basic system, but individuals can explicitly opt out of that and into a private insurance market (an option taken, as expected, mainly by upper-income Germans). In the United States, employer-provided and other nongroup health insurance can be thought of as *replacement private insurance* that individuals can voluntarily purchase to replace the implicit basic bundle of charity care and emergency services.[15] This is analogous to (disproportionately higher-income) people opting out of the public K–12 schools and instead paying out-of-pocket for private schools.

Replacement private insurance raises issues of its own. It results in lower public spending, but, depending on pricing institutions, it can also exacerbate adverse selection and market unraveling relative to add-on private insurance (Weyl and Veiga 2017)—potentially leading to a breakdown of risk pooling. Furthermore, when electing the replacement plan means losing the subsidy for basic coverage, this choice may result in inefficient crowding out of private spending. Even individuals who might be interested in additional health insurance beyond the basic bundle may decide against doing so because they wish to avoid losing a generous public subsidy (analogous to Peltzman 1973). In a US health care context, the fact that Medicaid provides long-term care insurance at no out-of-pocket cost—albeit only after household assets are drawn down substantially—is widely believed to have crowded out the provision of private long-term care insurance (Brown and Finkelstein 2008). The degree to which a basic bundle might crowd out private health insurance increases as the price of medical care increases.

This notion of a publicly guaranteed basic bundle alongside private supplemental coverage is of course controversial. Many people believe strongly in equal

[15] Medicaid is not exactly an implicit basic bundle. It offers free coverage—including being retroactive typically for 90 days—for those who are eligible. But eligibility is nonuniversal and individual eligibility varies greatly over time with changes in income and family structure. Given the hassles of enrollment, take-up is far from universal.

access to care regardless of income or ability to pay—not just a basic floor level of care provided to all. Allowing a top-up plan means that higher-income people are likely to have access to more care and better health outcomes. Some healthcare systems limit or block supplemental health insurance. A well-known example is Canada, which (by rule until 2005, and de facto today) disallows private insurance for services covered by its national Medicare-like system—though it allows private insurance for noncovered services, including prescription drugs. Similarly, many "Medicare for All" plans like that proposed by Senator Bernie Sanders disallow most private insurance. As with many other features of the social floor approach to universal coverage, there are likely to be important tradeoffs on which additional research would add great value.

Conclusion

Achieving meaningful universal coverage in the United States requires an explicit policy decision about is meant by that term. We argue that incremental expansions focused on addressing market failures in the current US system will propagate inefficiencies in our patchwork approach and will fail to facilitate the active policy decisions needed to achieve socially optimal coverage. By instead defining a basic bundle of valuable services that is publicly financed for all, while allowing individuals to "top up" by purchasing additional coverage, policymakers could both expand coverage to the uninsured and maintain incentives for innovation in a financially sustainable system.

Of course, there are important challenges to such a system redesign. Hard decisions would have to be made about tradeoffs among priorities for the allocation of scarce public resources (which are of course implicitly being rationed now)—opening up further potential for the politicization of medical decisions. Changing the functioning of the enormous US health care sector would be inherently disruptive—and perhaps particularly disruptive to the existing employer-sponsored insurance system, necessitating careful transition mechanisms.

To provide additional health care to the currently uninsured without substantially cutting back on care covered by existing public programs, such a system would also require a substantial increase in taxes, raising important questions about progressivity and deadweight loss. In this approach, private health insurance spending would be at least partially supplanted by government spending (with taxes rising commensurately, leaving similar take-home pay net of health care). There would be legitimate concerns about disruption to clinical relationships as provider networks realigned under new insurance coverage design, and legitimate fears about the government as monopsonist payer lowering the incentives for medical innovation by setting prices that reduce the ability of entrepreneurs to capture value, highlighting the importance of additional supplemental private insurance options.

Despite these challenges, few would argue that the current US health care system is serving everyone well. We are surely spending too much on the provision of health care that is delivering too little benefit to too few people. Reconceptualizing what we mean by universal coverage to ensure that public resources are devoted to care with high health benefit offers the opportunity to ensure universal access to innovative care in an affordable system.

■ *The authors thank the editors and Joe Newhouse for helpful feedback and suggestions, and Jason Furman and Matt Fiedler for sharing historical data on US uninsured rates.*

References

Abaluck, Jason, and Jonathan Gruber. 2011. "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program." *American Economic Review* 101 (4): 1180–210.

Anderson, Gerard F., Peter Hussey, and Varduhi Petrosyan. 2019. "It's Still the Prices, Stupid: Why the US Spends So Much on Health Care, and a Tribute to Uwe Reinhardt." *Health Affairs* 38 (1): 87–95.

Baicker, Katherine, and Amitabh Chandra. 2020. "What Values and Priorities Mean for Health Reform." *New England Journal of Medicine* 383 (15).

Baicker, Katherine, Amitabh Chandra. 2010. "Uncomfortable Arithmetic—Whom to Cover versus What to Cover." *New England Journal of Medicine* 362 (2): 95–97.

Baicker, Katherine, Amitabh Chandra, and Mark Shepard. 2023. "Replication data for: Achieving Universal Health Insurance Coverage in the United States: Addressing Market Failures or Providing a Social Floor?" American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/EXXXXXXV1.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein. 2015. "Behavioral Hazard in Health Insurance." *Quarterly Journal of Economics* 130 (4): 1623–67.

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor. 2017. "Choose to Lose: Health Plan Choices from a Menu with Dominated Option." *Quarterly Journal of Economics* 132 (3): 1319–72.

Blume-Kohout, Margaret E., and Neeraj Sood. 2013. "Market Size and Innovation: Effects of Medicare Part D on Pharmaceutical Research and Development." *Journal of Public Economics* 97: 327–36.

Brot-Goldberg, Zarek, Samantha Burn, Timothy Layton, and Boris Vabson. 2022. "Rationing Medicine through Bureaucracy: Authorization Restrictions in Medicare." Working Paper. https://zarekcb.github.io/PriorAuth_Web.pdf.

Brot-Goldberg, Zarek, Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad. 2017. "What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics." *Quarterly Journal of Economics* 132 (3): 1261–318.

Brown, Jason, Mark Duggan, Ilyana Kuziemko, and William Woolston. 2014. "How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program." *American Economic Review* 104: 10: 3335–64.

Brown, Jeffrey R., and Amy Finkelstein. 2008. "The Interaction of Public and Private Insurance: Medicaid and the Long-Term Care Insurance Market." *American Economic Review* 98 (3): 1083–102.

Buchanan, James M. 1975. "The Samaritan's Dilemma." In *Altruism, Morality and Economic Theory*, edited by Edmund S. Phelps, 71–86. New York: Russell Sage Foundation.

Chandra, Amitabh, and Craig Garthwaite. 2019. "Economic Principles for Medicare Reform." *Annals of the American Academy of Political and Social Science* 686 (1): 63–92.

Chandra, Amitabh, and Jonathan Skinner. 2012. "Technology Growth and Expenditure Growth in Health Care." *Journal of Economic Literature* 50 (3): 645–80.

Chandra, Amitabh, Evan Flack, and Ziad Obermeyer. 2021. "The Health Costs of Cost-Sharing." NBER Working Paper 28439.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight. 2010. "Patient Cost-Sharing and Hospitalization Offsets in the Elderly." *American Economic Review* 100 (1): 193–213.

Clemens, Jeffrey, and Joshua D. Gottlieb. 2014. "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?" *American Economic Review* 104 (4): 1320–49.

Clemens, Jeffrey, Joshua D. Gottlieb, and Jeffrey Hicks. 2021. "How Would Medicare for All Affect Health System Capacity? Evidence from Medicare for Some." *Tax Policy and the Economy* 35: 225–62.

Cutler, David M., and Dan P. Ly. 2011. "The (Paper)Work of Medicine: Understanding International Medical Costs." *Journal of Economic Perspectives* 25 (2): 3–25.

Curto, Vilsa, Liran Einav, Amy Finkelstein, Jonathan Levin, and Jay Bhattacharya. 2019. "Health Care Spending and Utilization in Public and Private Medicare." *American Economic Journal: Applied Economics* 11 (2): 302–32.

Dafny, Leemore S. 2010. "Are Health Insurance Markets Competitive?" *American Economic Review* 100 (4): 1399–431.

Dafny, Leemore, Mark Duggan, and Subramaniam Ramanarayanan. 2012. "Paying a Premium on Your Premium? Consolidation in the US Health Insurance Industry." *American Economic Review* 102 (2): 1161–85.

Domurat, Richard, Isaac Menashe, and Wesley Yin. 2021. "The Role of Behavioral Frictions in Health Insurance Marketplace Enrollment and Risk: Evidence from a Field Experiment." *American Economic Review* 111 (5): 1549–74.

Dorn, Stan, James C. Capretta, and Lanhee J. Chen. 2018. "Making Health Insurance Enrollment as Automatic as Possible (Part 1)." Health Affairs Blog, May 2. https://www.healthaffairs.org/do/10.1377/forefront.20180501.141197/full/.

Elhauge, Einer. 2011. "The Broccoli Test." *New York Times*, November 15. https://www.nytimes.com/2011/11/16/opinion/health-insurance-and-the-broccoli-test.html.

Einav, Liran, and Amy Finkelstein. 2011. "Selection in Insurance Markets: Theory and Empirics in Pictures." *Journal of Economic Perspectives* 25 (1): 115–38.

Ericson, Keith M., and Justin R. Sydnor. 2018. "Liquidity Constraints and the Value of Insurance." NBER Working Paper 24993.

Finkelstein, Amy. 2007. "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare." *Quarterly Journal of Economics* 122 (1): 1–37.

Finkelstein, Amy, Nathaniel Hendren, and Erzo F. P. Luttmer. 2019. "The Value of Medicaid: Interpreting Results from the Oregon Health Insurance Experiment." *Journal of Political Economy* 127 (6): 2836–74.

Finkelstein, Amy, Nathaniel Hendren, and Mark Shepard. 2019. "Subsidizing Health Insurance for Low-Income Adults: Evidence from Massachusetts." *American Economic Review* 109 (4): 1530–67.

Finkelstein, Amy, Neale Mahoney, and Matthew J. Notowidigdo. 2018. "What Does (Formal) Health Insurance Do, and for Whom?" *Annual Review of Economics* 10: 261–86.

Furman, Jason and Matt Fiedler. 2014. "2014 Has Seen Largest Coverage Gains in Four Decades, Putting the Uninsured Rate at or Near Historic Lows." White House Council of Economic Advisers blog, December 18. https://obamawhitehouse.archives.gov/blog/2014/12/18/2014-has-seen-largest-coverage-gains-four-decades-putting-uninsured-rate-or-near-his.

Garthwaite, Craig, Tal Gross, and Matthew J. Notowidigdo. 2018. "Hospitals as Insurers of Last Resort." *American Economic Journal: Applied Economics* 10 (1), 1–39.

Geruso, Michael, and Timothy J. Layton. 2017. "Selection in Health Insurance Markets and Its Policy Remedies." *Journal of Economic Perspectives* 31 (4): 23–50.

Geruso, Michael, and Timothy Layton. 2020. "Upcoding: Evidence from Medicare on Squishy Risk Adjustment." *Journal of Political Economy* 128 (3): 984–1026.

Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *American Economic Review* 103 (7): 2643–82.

Holgash, Kayla and Martha Heberlein. 2019. "Physician Acceptance of New Medicaid Patients." *Medicaid and CHIP Payment and Access Commission*, January 24. https://www.macpac.gov/wp-content/uploads/2019/01/Physician-Acceptance-of-New-Medicaid-Patients.pdf.

Kaiser Family Foundation. 2021a. "Distribution of Eligibility for ACA Health Coverage among the Remaining Uninsured." https://www.kff.org/health-reform/state-indicator/distribution-of-eligibility-for-aca-coverage-among-the-remaining-uninsured.

Kaiser Family Foundation. 2021b. "Health Insurance Coverage of the Total Population." https://www.kff.org/other/state-indicator/total-population.

Madrian, Brigitte C. 1994. "Employment-Based Health Insurance and Job Mobility: Is There Evidence of Job-Lock?" *Quarterly Journal of Economics* 109 (1): 27–54.

Mahoney, Neale. 2015. "Bankruptcy as Implicit Health Insurance." *American Economic Review* 105 (2): 710–46.

Mahoney, Neale, and E. Glen Weyl. 2017. "Imperfect Competition in Selection Markets." *Review of Economics and Statistics* 99 (4): 637–51.

McIntyre, Adrianna, Mark Shepard, and Myles Wagner. 2021. "Can Automatic Retention Improve Health Insurance Market Outcomes?" *AEA Papers and Proceedings* 111: 560–66.

McWilliams, J. Michael, Laura A. Hatfield, Bruce E. Landon, Pasha Hamed, & Michael E. Chernew. 2018. "Medicare Spending after 3 Years of the Medicare Shared Savings Program." *New England Journal of Medicine* 379 (12), 1139–49.

Newhouse, Joseph P. 1993. *Free for All? Lessons from the RAND Health Insurance Experiment.* Cambridge, MA: Harvard University Press.

Newhouse, Joseph P., and Thomas G. McGuire. 2014. "How Successful is Medicare Advantage?" *Milbank Quarterly* 92 (2): 351–94.

OECD. 2022. "Social Protection: Total Public and Primary Voluntary Health Insurance." OECD Health Statistics. https://stats.oecd.org/Index.aspx?QueryId=30137.

Peltzman, Sam. 1973. "The Effect of Government Subsidies-in-Kind on Private Expenditures: The Case of Higher Education." *Journal of Political Economy* 81 (1): 1–27.

Rae, Matthew, Cynthia Cox, Gary Claxton, Daniel McDermott, and Anthony Damico. 2021. "How the American Rescue Plan Act Affects Subsidies for Marketplace Shoppers and People Who Are Uninsured." Kaiser Family Foundation, March 25. https://www.kff.org/health-reform/issue-brief/how-the-american-rescue-plan-act-affects-subsidies-for-marketplace-shoppers-and-people-who-are-uninsured/.

Shepard, Mark, and Myles Wagner. 2022. "Reducing Ordeals through Automatic Enrollment: Evidence from a Subsidized Health Insurance Exchange." Working Paper. https://scholar.harvard.edu/files/mshepard/files/shepard_wagner_autoenrollment.pdf.

Shepard, Mark, Katherine Baicker, and Jonathan Skinner. 2020. "Does One Medicare Fit All? The Economics of Uniform Health Insurance Benefits." *Tax Policy and the Economy* 34: 1–41.

Sommers, Benjamin D., Atul A. Gawande, and Katherine Baicker. 2017. "Health Insurance Coverage and Health—What the Recent Evidence Tells Us." *New England Journal of Medicine* 377 (6): 586–93.

Spinnewijn, Johannes. 2017. "Heterogeneity, Demand for Insurance, and Adverse Selection." *American Economic Journal: Economic Policy* 9 (1): 308–43.

Starc, Amanda. 2014. "Insurer Pricing and Consumer Welfare: Evidence from Medigap." *RAND Journal of Economics*: 45 (1), 198–220.

Tax Policy Center. 2022. "How Does the Tax Exclusion for Employer-Sponsored Health Insurance Work?" https://www.taxpolicycenter.org/briefing-book/how-does-tax-exclusion-employer-sponsored-health-insurance-work.

Tebaldi, Pietro. 2022. "Estimating Equilibrium in Health Insurance Exchanges: Price Competition and Subsidy Design under the ACA." NBER Working Paper 29869.

Tolbert, Jennifer, Kendal Orgera, and Anthony Damico. 2019. "Key Facts about the Uninsured Population." Kaiser Family Foundation, Issue Brief. https://files.kff.org/attachment//fact-sheet-key-facts-about-the-uninsured-population.

US Centers for Medicare and Medicaid Services. 2022. "National Health Expenditure Data." https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData.

Weisbrod, Burton A. 1991. "The Health Care Quadrilemma: An Essay on Technological Change, Insurance, Quality of Care, and Cost Containment." *Journal of Economic Literature* 29 (2): 523–52.

Weyl, E. Glen, and Andre Veiga. 2017. "Pricing Institutions and the Welfare Cost of Adverse Selection." *American Economic Journal: Microeconomics* 9 (2): 139–48.

Wright, Bill J., Ginny Garcia-Alexander, Margarette A. Weller, and Katherine Baicker. 2017. "Low-Cost Behavioral Nudges Increase Medicaid Take-Up Among Eligible Residents of Oregon." *Health Affairs* 36 (5): 838–45. doi: 10.1377/hlthaff.2016.1325.

Journal of Economic Perspectives—Volume 37, Number 2—Spring 2023—Pages 123–152

The Prices in the Crises: What We Are Learning from 20 Years of Health Insurance in Low- and Middle-Income Countries

Jishnu Das and Quy-Toan Do

By the late 1990s, health systems in most low-income countries provided subsidized care at public clinics funded through general taxation. Although there was wide variation in how often public clinics were used for primary care (ranging from very little in India to almost exclusively in many Latin American countries), for hospitalization and inpatient care the public sector consistently accounted for 50–80 percent of health care provision across multiple countries, a share that has remained remarkably stable over time (World Health Organization 2020; Grépin 2016). Perhaps surprisingly, this wide availability of subsidized public clinics coexisted with high out-of-pocket expenditures. Indeed, two decades ago out-of-pocket expenses amounted to 50 percent of total health expenditure in low-income countries, with households frequently unable to insure themselves against large health shocks and the related loss of income (Gertler and Gruber 2002). Summarizing the situation at the turn of the twenty-first century, Pauly et al. (2006) observed: "Virtually every developing country with a functioning government uses publicly-funded and managed systems for third-party payment for medical care. . . . [but] in many developing countries, this system has failed to provide adequate financial protection for its citizens and adequate access to care. The gap shows up in the form of private out-of-pocket spending for services that 'universal insurance' cannot or does not supply."

■ *Jishnu Das is Professor, McCourt School of Public Policy & Walsh School of Foreign Service, Georgetown University, Washington, DC. He is also a Faculty Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Quy-Toan Do is a Lead Economist, Development Research Group, World Bank, Washington, DC. Their email addresses are Jishnu.das@georgetown.edu and qdo@worldbank.org.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.123.

In response, governments in many low- and middle-income countries have moved towards some form of dedicated health insurance, which they believed would provide the instruments needed to address problems related to the financing and provision of healthcare. Insurance premiums would provide a dedicated source of funding for healthcare, while simultaneously reducing out-of-pocket expenditures and providing much-needed financial protection for citizens. Furthermore, governments postulated that insurance products would improve health outcomes by altering the behavior of patients and providers. By subsidizing the cost of visiting private sector providers, health insurance would expand patient choice and increase visits to participating private sector clinics, where the quality was thought to be higher. By redesigning how public and private providers were reimbursed, health insurance would also increase the incentives to provide higher-quality care, especially in the public sector. The transition in how providers were reimbursed, referred to as the shift from passive to "strategic purchasing" in the health literature (World Health Organization 2000; Hanson et al. 2019), would then improve the quality and cost effectiveness of health service delivery (Londoño and Frenk 1997).

In this essay, we evaluate the experience with health insurance in low- and middle-income countries over the last 20 years, where health insurance is provided by a specific scheme that is additional to the subsidized care available through public clinics. We start by documenting the transition to health insurance, drawing on data from 100 Demographic and Health Surveys conducted across 62 countries between 1989 and 2019. Using country-level examples and a review of existing evaluations, we describe how governments have structured their health insurance schemes and how these schemes have functioned in practice. Our conclusion from this review is that health insurance schemes have successfully increased financial protection and utilization, but there is little evidence (yet) of improvements in health outcomes. Further, there has been little demand for health insurance among households, even when it is heavily subsidized.

We then discuss—and rule out—the possibility that health insurance has not improved health outcomes because of systemic constraints in the delivery of healthcare. Instead, we believe that health insurance triggered behavioral responses among providers that have systematically undermined the objectives of insurance schemes. These responses have led to prices that are higher than those mandated under the program, an increase in unnecessary care, and new sources of uncertainty as insured patients do not know whether the insurance will be honored or whether they will be correctly treated.

While such behavioral responses—sometimes called "provider moral hazard"—are also a concern in high-income countries, health insurance schemes in these countries typically work with a variety of institutions that seek to curb these forms of physician excess, including review boards, professional norms, public reporting, and punitive enforcement through courts.[1] However, nonprice incentive systems in

[1] For examples of provider moral hazard in a US context, see Gruber and Owings (1996) on the use of C-sections, Baker (2010) on how doctors prescribe more magnetic resonance imaging after buying an

low- and middle-income countries are poorly developed and have not been used successfully in combination with health insurance schemes, which can create social tensions in which patients take matters into their own hands; as one example, there were an estimated 17,000 attacks and agitations against doctors in China in 2010 alone (Tussing, Wang, and Wang 2014)

We conclude that the first 20 years of experience with health insurance schemes in low- and middle-income countries shows that it is impossible to divorce the quality of care that health insurance offers from the financial protection it affords. Insurance is realized as a subsidy to patients when they seek care, and the flip side of that subsidy is a payment to providers. But providers respond to the financial incentives created by how those payments are structured and therefore the value of the insurance is critically tied to the nature of these responses. The lack of nonprice mechanisms combined with the difficulties of structuring price incentives for appropriate care imply that health insurance could prove to be a very expensive tool for improving financial protection that actually *lowers* the quality of health care in low- and middle-income countries.

How are Health Insurance Schemes Structured in Low- and Middle-Income Countries?

We start by mapping the overall coverage of health insurance across low- and middle-income countries using data from the Demographic and Health Surveys, which are nationally representative health-related surveys that target women aged 15–49 and their children under the age of five in low- and middle-income countries, with a traditional focus on maternal and child health. Questions on health insurance coverage were sporadically included in the late 1990s for some countries, like Bolivia, Jordan, Peru, and Turkey, and a standard question (v481) was introduced in 2003. Of 426 surveys ever fielded, 100 surveys in 62 countries have health insurance data available as of April 2022. We compiled all the health insurance questions to create a single database of just under five million observations, representative of countries with a total population of about 3.5 billion people (for additional details, see online Appendix A). When combined with OECD health insurance data (Scheil-Adlung 2014), our DHS-based data allow us to present a unified picture of health insurance coverage and its correlates across multiple countries, adding to previous studies that have focused on single or small groups of countries (for example, Amu et al. 2022; Barasa et al. 2021; Wang, Temsah, and Mallick 2014; National Population Commission and ICF International 2014).

For low- and middle-income countries where data were available from either the Demographic and Health Surveys or the OECD, close to half of the total population now reports that they have health insurance, as shown in Table 1. Coverage

MRI machine for their clinic, and Clemens and Gottlieb (2014) on how treatment choices depend on reimbursement rates.

Table 1
Health Insurance Coverage across Countries and over Time

Country (1)	*First measured coverage rate (%), Year (2)*	*Latest measured coverage rate (%), Year (3)*
Afghanistan		0.11 (2015)
Albania	23.31 (2008)	29.88 (2017)
Angola		5.32 (2015)
Armenia	0.76 (2010)	7.82 (2015)
Azerbaijan		1.17 (2006)
Bangladesh		0.18 (2017)
Benin	1.4 (2011)	1.09 (2017)
Bolivia	20.07 (1989)	26.57 (2008)
Burkina Faso		0.51 (2010)
Burundi	13.63 (2010)	23.18 (2016)
Cambodia	13.61 (2010)	14.57 (2014)
Cameroon	1.64 (2011)	2.17 (2018)
Chad		0.3 (2014)
Comoros		5.4 (2012)
Congo		2.01 (2011)
Congo, Democratic Rep.		4.13 (2013)
Cote d'Ivoire		2.82 (2011)
Dominican Republic		56.5 (2013)
Egypt	11.78 (2005)	9.88 (2014)
Ethiopia	0.89 (2010)	4.2 (2016)
Gabon		54.03 (2012)
Gambia	2.13 (2013)	2.17 (2019)
Ghana	**41.7 (2008)**	**65.95 (2014)**
Guam		12.5 (2014)
Guinea		1.44 (2018)
Guyana		16.33 (2009)
Haiti	1.59 (2005)	2.24 (2016)
Honduras	8.19 (2005)	7.66 (2011)
India	**5.9 (2005)**	**17.86 (2015)**
Indonesia	**40.71 (2012)**	**61.36 (2017)**
Jordan		70.32 (2017)
Kenya	7.15 (2008)	7.21 (2014)
Kyrgyzstan		88.28 (2012)
Lesotho	8.74 (2009)	1.65 (2014)
Liberia	3.58 (2013)	3.69 (2019)
Madagascar		1.88 (2008)
Malawi		1.37 (2015)
Maldives		4.86 (2016)
Mali	2.51 (2012)	4.76 (2018)
Moldova		54.91 (2005)
Mozambique	3.08 (2011)	2.97 (2015)
Myanmar		0.96 (2015)
Namibia	**15.77 (2006)**	**17.73 (2013)**
Niger		3.58 (2012)
Nigeria	1.79 (2008)	2.67 (2018)
Pakistan		2.57 (2017)
Papua New Guinea		3.93 (2016)
Peru	26.57 (2003)	58.05 (2012)
Rwanda	**41.04 (2005)**	**83.22 (2019)**
Sao Tome and Principe		1.8 (2008)
Senegal	5.11 (2010)	5.7 (2016)
Sierra Leone	2.12 (2008)	3.47 (2019)
South Africa		6.31 (2016)
Swaziland		5.31 (2006)
Tanzania	4.9 (2009)	8.13 (2015)
Togo		4.55 (2013)
Turkey	**56.23 (1998)**	**88.58 (2013)**
Uganda	1.61 (2011)	1.29 (2016)
Yemen		1.77 (2013)
Zambia	7.16 (2007)	1.86 (2018)
Zimbabwe	8.77 (2005)	12.21 (2015)

Source: See online Appendix B for documentation of the data used for this analysis.
Notes: The table reports health insurance coverage rates using the Demographic and Health Surveys dataset for surveys from the year 2000 or later. When data is available for only one year, no information is reported in column 2. Countries in **bold** are those referred to in the text.

is near-universal for several countries in East and Central Asia and Latin America and middling in India and some sub-Saharan countries (Namibia, Rwanda, and Ghana). Other countries in sub-Saharan Africa and South Asia are still below 5 percent coverage. Interestingly, most of the growth in health insurance coverage is relatively recent: in Rwanda, an increase from 41 percent in 2005 to 83 percent by 2019; in Turkey, from 56 percent in 1998 to 88 percent in 2013; and in Indonesia, from 40 percent in 2012 to 61 percent in 2017. In historical context, this expansion is quite rapid, given that European countries typically took 60–70 years to expand health insurance coverage from 10 to 20 percent around the turn of the twentieth century to above 75 percent in 1975 (Tanzi and Schuknecht 2000; Ortiz-Ospina and Roser 2017).

The architecture of these health insurance schemes encompasses two features: how this coverage is financed and how providers are reimbursed.[2] We document that health insurance schemes that have emerged in low- and middle-income countries during the last two decades are similar when it comes to financing, but differ substantially in how they compensate physicians.

Financing

To provide a concrete example, we anchor our discussion around the federal health insurance scheme that India introduced in 2008, called the Rashtriya Swasthya Bima Yojana or RSBY ("National Health Insurance Scheme"). Before 2008, public health care for Indians was free or subsidized, but most people still chose to visit private sector facilities where, lacking health insurance, they were charged market-determined prices. Consequently, out-of-pocket expenditures were the main source of health financing, and rising healthcare costs were identified as a key factor driving households into debt. The government subsequently introduced the RSBY, under which households below the poverty line could enroll into the scheme by paying the nominal amount of ₹30 in Indian rupees (about $0.40 in US dollars) for a maximum of five members per household (Palacios, Das, and Sun 2011). Once enrolled, they could use their insurance for a range of inpatient services free of charge in participating private or public hospitals up to an annual limit of ₹30,000 per household, subsequently increased to ₹500,000 ($6,500) in 2018, when the scheme was also renamed PM-JAY (The Prime Minister's People's Health Scheme). To administer the scheme, every participating state solicited bids annually from insurance companies in the form of a per-household premium, and winning companies were paid according to the number of households that they enrolled by the federal and state governments in a 75/25 cost-sharing agreement. The per-household premium that emerged through the bidding process ranged from ₹500 to ₹600 ($6.60 to $8) in the initial years. Insurance companies contracted independently with participating hospitals and reimbursed them using administrative prices that were determined by the state.

India's RSBY program is similar to health insurance schemes that have emerged in other low- and middle-income countries in two important ways. First, by the time RSBY was implemented, it was clear from the experience of other countries as well as pilot insurance schemes within India that the demand for unsubsidized

[2] A third feature of health insurance schemes is what is covered and to what extent. A rich tradition going back to Spence and Zeckhauser (1971) discusses the optimal design of coverage schemes to tradeoff incentives against insurance. A large empirical literature in the United States exploits features of these schemes to better understand the impact of deductibles and coverage limits. While the analysis of coverage determination and its consequences constitutes a research agenda on its own, we do not pursue this line of inquiry here beyond acknowledging the wide heterogeneity across countries in coverage scale and scope. This partly reflects the state of the literature—we have not found studies that examine the consequences of coverage schemes on the outcomes we study, which reflects the early stages of health insurance in these countries. It also reflects our understanding that the questions of financing and reimbursement discussed here would remain relevant if policymakers were to consider expanding the scope of insurance coverage.

insurance among poor households was very low. Therefore, India's government subsidized the premium from the very beginning using funds collected through a variety of direct and indirect taxes. This reliance on taxes rather than premiums to fund health insurance is now common in most low- and middle-income countries. For instance, in 1993, Colombia's Law 100 led to universal health insurance financed through mandatory payroll contributions in a contributory regime and general taxation in a subsidized regime, with the beneficiaries in the latter identified through a proxy-means test (Escobar et al. 2009). Ghana's health insurance scheme is financed through a combination of value-added taxes (70 percent) and social security taxes (23 percent), with premiums accounting for another 5 percent (Blanchet, Fink, and Osei-Akoto 2012). In Vietnam, the government tried at first to collect funds through premiums, but has since moved to financing based on general taxation or mandatory contributions (Somanathan et al. 2014). Kenya, where the National Health Insurance Fund established in 1966 was designed to provide coverage for formal sector workers financed by mandatory contributions, is now moving towards expanding coverage through a health insurance subsidy program for the poor (Barasa et al. 2018). In fact, across low- and middle-income countries, voluntary purchases of private health insurance in 2012 covered less than 1 percent of the population in 49 of 138 countries and 1–5 percent in another 39 countries (Drechsler and Jütting 2005; Pettigrew and Mathauer 2016).[3]

Second, India's RSBY was not new insurance. Instead, it was layered on top of existing access to free or highly subsidized care through public clinics, themselves financed through general taxation. This new layer was designed to improve healthcare by reducing the cost of using private facilities, where quality was believed to be higher, and by moving to a system where money follows the patient—with the hope of providing incentives for quality improvements in both public and private health facilities. Both of these features—health insurance layered on existing networks of subsidized public care and health insurance used as a mechanism for improving quality—are again common across low- and middle-income countries. Of course, the relative importance of these two mechanisms depends on the context. In Latin America, for instance, the share of the private sector is smaller, and insurance schemes tried to address differences in quality between clinics run by the social

[3] The ubiquity of public subsidies for financing raises the important question of whether coverage differs by household characteristics. Household wealth and education are two widely-used markers of socioeconomic status in low- and middle-income countries, the former measured by an index of asset ownership and the latter by completion of different levels of schooling. We estimated correlations of insurance coverage and wealth, as well as insurance coverage and education for 61 countries using the latest available round of Demographic and Health Surveys data, and plot these coefficients in online Appendix Figure A1, with data and details available in the line index. In most countries, health insurance coverage is regressive, with higher coverage for those with greater education and more wealth. This pattern reflects both household demand for insurance and policy choices. In some countries, particularly those in Latin America, health insurance has always been available for government employees and was then extended to those in the formal sector with mandatory contributions, which explains the regressive pattern. In contrast, India's health insurance scheme is targeted to those below the poverty line, and the correlation of coverage with education and wealth is close to zero in the DHS data.

security system (accessible only to those employed in the public sector) and those run by the public health system, rather than between private and public care.

Provider Reimbursements

Interestingly, the convergence that we see across countries in how health insurance is financed breaks down when we look at how providers are reimbursed. The RESYST ("Resilient and Responsive health systems") study from the London School of Hygiene and Tropical Medicine documents 19 different purchasing mechanisms in 10 low- and middle-income countries (Hanson et al. 2019). These include, among other things, fee-for-service (the hospital is reimbursed for each service given as part of the stay), capitation (physicians agree to a fixed amount per patient under their care for a given duration), diagnostic-related groups (bundle all goods and services for one hospitalization episode into a single price depending on patient characteristics), line-item budgeting (health ministry pays for each item according to a budget), and global budgeting (a hospital signs a contract for a sum over a period to cater to the population in its catchment), used either by themselves or in combinations. Prices may be set through individual negotiations, group negotiations or administratively. Administrative prices themselves vary in their degree of sophistication, ranging from average accounting costs of procedures to more sophisticated, risk-adjusted marginal cost pricing.[4]

Reimbursement rates often seem to involve political processes and large discontinuous jumps. In Kenya, prices paid to hospitals for surgeries were revised upward by 50 to 100 percent in some cases in 2016 (Barasa et al. 2018). In Vietnam, hospitals are reimbursed according to a predetermined fee-for-service schedule. Tien et. al (2011) note that, at the time of their article, the price of services on the original list had not been updated since 1995, although an additional 992 services were added in 2006. In Ecuador, prices have not been updated since 2012, even in nominal terms. In Colombia, prices for essential services are determined administratively rather than by the market, but the government is required to cover procedures outside the essential group if mandated by a court. From 2005 to 2010, the reimbursements for these additional procedures increased from 0.1 to 2.4 trillion pesos (about $607 million in US dollars), leading to a financial emergency in 2011, following which "reference pricing" was introduced for medicines to reduce the fiscal burden (Romero 2014; Inter-American Development Bank 2015; Giedion and Uribe 2009).

Thus, while the paths that countries have taken to reach this point have been very different, low- and middle-income countries have now arrived at a point where most subsidize their health insurance schemes and recognize that reimbursements for healthcare providers are integral to the scheme, even if there is no consensus on how these reimbursements should be structured. We stress that multiple features of these schemes are different from what economists typically think of as a textbook insurance system. First, these schemes complement preexisting and heavily

[4] Barber, Lorenzoni, and Ong (2019) document similar variation in remuneration schemes for providers in their study of OECD countries, along with Thailand and Malaysia.

subsidized public systems, which paradoxically are not referred to as health insurance even though the financing is identical. Second, an important implication of the fact that the premiums are now publicly financed is that the issue of adverse selection in health insurance is less relevant. Private insurance companies must make positive profits from the insurance product and thus are not viable if only the sickest patients subscribe. Such concerns do not apply with publicly subsidized insurance, as the government can always cover any financing gap in the insurance scheme through taxes. In the Indian RSBY for instance, for a given premium, the scheme is cheaper for the government when only the sickest patients participate. Third, the health insurance schemes often cover preventive care, which is not risky and therefore is unrelated to the "insurance" part of health insurance. These differences lead us to believe that it may be better to think of health insurance schemes in low- and middle-income countries as contracting mechanisms that modify reimbursement schemes for public providers and expand existing networks to private providers. Nevertheless, in the interest of maintaining current convention, we continue to label these schemes as "health insurance." Our caution is that intuition about "insurance" can be misleading in this context, to the extent that it leads us down the path of demand-side failures and adverse selection rather than supply-side questions of contracting and provider incentives.

How Have Health Insurance Schemes Performed?

To study the effects of the expansion in health insurance, we now review the research that examines the link between health insurance coverage, financial protection, utilization, and ultimately health outcomes. We then discuss whether these findings are consistent with what we know of the demand for health insurance among households.

The main methodological challenge that studies of health insurance face is that the demand for health insurance and the benefits that accrue to households are likely to be correlated with underlying health status: patients with higher expected demand for health care should also be the ones most likely to take-up the insurance product, use it, and benefit from it (Wagstaff et al. 2016; Thornton 2021; Spenkuch 2012). The studies that we discuss all take exceptional care in addressing this fundamental identification challenge, either by using randomized experiments that incentivize take-up or by exploiting natural experiments that generate variation in eligibility across space and over time. As specific examples, King et al. (2009) choose a random sample of communities in Mexico to implement a health insurance scheme a year earlier than expected, Fink et al. (2013) randomize the rollout of a community-based social health insurance scheme in Burkina-Faso, and Sood and Wagner (2018) leverage a staggered roll-out of a social insurance scheme in the Indian state of Karnataka.[5]

[5] We exclude several observational studies that use difference-in-differences approaches, as recent work highlights the importance of weighting and functional form in such studies and the literature precedes

Financial Protection

Does health insurance achieve its primary stated goal of reducing household out-of-pocket expenses? The answer is an unambiguous yes: out-of-pocket healthcare expenditures decline with health insurance in low- and middle-income countries, as does the variability of such expenditures. This robust result holds across studies that use different measures of financial protection and across countries where insurance products differ in terms of coverage and benefits.

Bauhoff, Hotchkiss, and Smith (2011) document that Georgia's Medical Insurance Program for the poor led to a 50 percent decline in out-of-pocket expenses over a 30-day recall period, which the authors attribute to a 20 percent lower likelihood of incurring any health expenditure at all. The declines are quite remarkable as the program does not cover the cost of medicines, which account for 50 percent of total health care expenditures. Powell-Jackson et al. (2014) report a 27 percent decline in total health expenditures for the insured in Ghana, again using a four-week recall period. The insurance scheme in Ghana covers basic care (preventive care and medicines) as well as secondary care procedures, all of which add to the free care already available in public facilities. King et al. (2009) study *Seguro Popular,* Mexico's universal coverage insurance scheme, which covers 266 health interventions, 312 medicines, and a federal fund for catastrophic health expenditures for certain diseases. Using a ten-month recall period, the authors find an 85 percent decline in out-of-pocket expenditures and a 75 percent lower probability of catastrophic expenditures, the latter defined as cases where out-of-pocket expenditures exceeded 30 percent of a subsistence income level.

Defining catastrophic expenditures as a fraction of household income rather than a standard subsistence income level yields similar results. Celhay et al. (2019) report a 15 percent decline in the likelihood of such events in the Philippines when catastrophic events are defined as those where expenditures exceeded 10 percent of income; Fink et al. (2013) find a 30 percent decreased likelihood when the cutoff is defined at 5 percent in Burkina Faso. Yet another metric, adopted by Levine, Polimeni, and Ramage (2016), focuses on absolute thresholds of annual expenditures above US$250 or more than US$100 paid for a single event as well as instances of indebtedness due to health care payment obligations. They find a 20 percent

these econometric developments, making it harder for us to assess the validity of the estimates (Roth et al. 2022). Unfortunately, this includes results concerning India's RSBY program and China's New Cooperative Medical Scheme. As these are important programs and cover close to half of the world's population in low- and middle-income countries, we mention the findings from relevant studies here. For India, Karan, Yip, and Mahal (2017) show that the RSBY had zero impact on financial protection, a null result that has also been shown for state-level samples from the state of Chattisgarh (Garg, Bebarta, and Tripathi 2020) and for the three southern India states of Andhra Pradesh, Karnataka, and Tamil Nadu (Garg, Chowdhury, and Sundararaman 2019). For China, several studies have looked at the impact of the New Cooperative Medical Scheme that provides health insurance to the rural population. There is little consensus in this literature and results seem sensitive to the exact specification, controls, and estimation techniques used. As Liang et al. (2012) highlight in their systematic review: "[I]ndividual studies indicated that NCMS had positive, negative, or no effect on health outcomes and/or the incidence of catastrophic health payments."

decline in the likelihood of catastrophic out-of-pocket health expenditures. The authors attribute this to the free health services and drugs made available by the insurance scheme at public facilities, suggesting that the public sector is used as an alternative to the private sector for large expenditures, especially when these involve taking on debt.

Utilization and Health Outcomes

Access to health insurance also seems to increase utilization for a variety of health services. Evidence of increased utilization has been documented for preventive care (in Colombia, Camacho and Conover 2013; in India, Malani et al. 2021; in Peru, Bernal, Carpio, and Klein 2017), outpatient visits for acute or chronic diseases (in Nigeria, Fitzpatrick and Thornton 2019), and inpatient visits including surgeries (in India, Sood and Wagner 2018; Malani et al. 2021).[6] Consistent with the idea that health insurance can lead patients to choose higher-quality facilities, Thornton et al.'s (2010) study in Nicaragua and Levine, Polimeni, and Ramage's (2016) study in Cambodia do find that patients substituted away from public and non-networked private facilities towards networked hospitals. Moreover, Powell-Jackson et al. (2014), using data from Ghana, Sood and Wagner (2018) from India, and Celhay et al. (2019) from Argentina document that these kinds of substitutions can increase the quality of care. No study to date looks at the impact of health insurance using facility-specific measures of quality; thus, it is possible that the studies that do not find any change in aggregate utilization are still missing changes on this margin.

In contrast to the widespread evidence of increased health care utilization, the impacts of health insurance on health outcomes have been mixed at best. Even when a comprehensive benefit package is offered, as in Mexico with the *Seguro Popular* program, King et al. (2009) fail to detect differences in health outcomes as measured by nine different self-assessments.[7] This basic result of zero to small impact resonates across a number of studies. Levine, Polimeni, and Ramage (2016) report zero impacts from Cambodia; Bauhoff, Hotchkiss, and Smith (2011) report zero results from Georgia; and Fink et al. (2013) report the same from Burkina Faso. Similarly, Powell-Jackson et al. (2014) do not find any health impacts in Ghana, and Miller, Pinto, and Vera-Hernández (2013) did not find significant effects of insurance on health outcomes in Colombia, whether these are self-reported health assessments, symptoms (like fever, cough, diarrhea,

[6] A few studies do not find that health insurance increases utilization: King et al. (2009) in a study of Mexico; Raza et al. (2016) in rural India; and Bauhoff, Hotchkiss, and Smith (2011) in Georgia. Of course, these findings would also explain why health insurance in these countries does not improve health outcomes.

[7] King et al.'s (2009) assessment came ten months after the inception of *Seguro Popular*, which may be too short a time period for impacts on health outcomes to emerge. Although their study cannot exploit the original randomization, Cohen and Dechezleprêtre (2022) find that three to four years after the insurance scheme was introduced, mortality rates appear to have reduced.

or blood pressure), or summary outcomes such as weight (including low birthweight), height, and mortality.

In the handful of studies that do find a positive causal impact on health outcomes, it is usually associated with the increased utilization of higher quality, especially preventive care. Sood and Wagner (2018) investigate the impact of a social health insurance program for the poor in the Indian state of Karnataka and find that it reduced mortality from heart conditions and cancer; they argue that insurance coverage led patients to seek early diagnosis. Likewise, in Camacho and Conover (2013), Balsa and Triunfo (2021), and Celhay et al. (2019), health insurance schemes in Colombia, Uruguay, and Mexico, respectively, are found to lead to reduced infant mortality, which is mostly attributable to increased use of preventive prenatal care.

The Demand for Health Insurance

The result that health insurance increases utilization but not outcomes is troubling, because it suggests that spending is increasing without anything to show for it. That drastic conclusion, though, must be modified both because the statistical power to detect impacts for relatively rare outcomes like mortality requires very large sample sizes that are often unavailable, and because studies to date may not have measured the health outcomes that improved, such as mental health. An alternative approach is to focus on demand and ask whether households are willing to purchase health insurance in the first place. Interestingly, and to the extent that demand is an appropriate measure of value in this case, households appear not to value health insurance very highly.

Several studies recover the demand for health insurance by experimentally varying financial and nonfinancial incentives for enrollment. The studies show that insurance take-up is low in the first place and large financial subsidies are required to increase enrollments significantly. Only in studies that offer government health insurance for free and on a continuing basis do we see significant increases in enrollments. However, even these steep discounts fail to achieve 100 percent enrollment rates, and enrollment drops as soon as subsidies cease. In an early example, Thornton et al. (2010) conducted a randomized controlled trial in Nicaragua that incentivizes informal sector workers to enroll in the Nicaraguan Social Security Institute's health insurance program, varying information on the program, the premium subsidy, and the type of insurance sales agents. Thornton et al. (2010) reported a 20 percent increase in enrollment when the premium is fully subsidized, followed by a 90 percent dropout rate at the end of the subsidy period. In the Philippines, Capuno et al. (2016) encouraged enrollment in a social health insurance program by experimentally varying information on the scheme, a financial incentive (up to a 50 percent premium subsidy), and administrative assistance. They find a 3 percentage-point increase in enrollment when a subsidy of 50 percent was offered, from a baseline rate of 8 percent. Subsequent interventions have been similar in spirit, with some variations. For instance, Banerjee et al. (2021) offer full and partial subsidies for insurance premiums in

Indonesia to examine whether household behavior changes when the price is a small positive amount rather than an exact zero. They find that a full premium subsidy increases enrollment to 19 percent from a baseline of 8 percent with retention rates 4.6 and 3.9 percentage points higher in the subsidized group at three and eight months after the subsidies ended. Other studies find similar results: Levine, Polimeni, and Ramage (2016) in rural Cambodia; Wagstaff et al. (2016) for informal sector workers in Vietnam; and Banerjee, Duflo, and Hornbeck (2014) among microfinance borrowers in India.

All these studies suggest that households do not value health insurance highly, which is consistent with the lack of evidence on the link between health insurance and health outcomes. The wrinkle that remains is that this finding is not automatically consistent with the evidence that health insurance improves financial protection. If *other* constraints hold back demand, it becomes difficult to interpret the price-elasticity as a measure of value. Poor information is a first possibility, but interventions that include an information component typically find little impact on health insurance take-up (Capuno et al. 2016; Wagstaff et al. 2016; Thornton et al. 2010; Banerjee et al. 2021; Malani et al. 2021; Das et al. 2016). A second possibility is that there are nonprice barriers to take-up, such as administrative burdens arising from eligibility requirements. Interventions that offer administrative assistance with enrolling in health insurance indeed find that take-up increases, with an effect size equivalent to that of premium subsidies for six months (Capuno et al. 2016; Thornton et al. 2010; Banerjee et al. 2021; Malani et al. 2021). This effect is sufficiently large that a simple explanation rooted in the opportunity cost of time is unlikely. As of yet, no consensus has emerged on the extent to which low demand reflects low expected value rather than administrative burdens; this remains very much at the frontier of the research on health insurance.

Why Has Health Insurance Not Improved Health Outcomes: Supply Constraints

The most obvious explanation for why health insurance has not improved health outcomes is that there is little capacity for quality improvement to begin with; for example, doctors are overworked, do not have the right equipment, and may not have the right skills. In this section we will show that, in contrast to this view—and despite severe deficits in quality—there is in fact considerable room for improvement. In the next section, we then consider what we view as a more likely possibility—that government health insurance in low- and middle-income countries has had an adverse effect on provider incentives, ultimately undermining the objectives of these schemes in a way that can *worsen* the quality of healthcare.

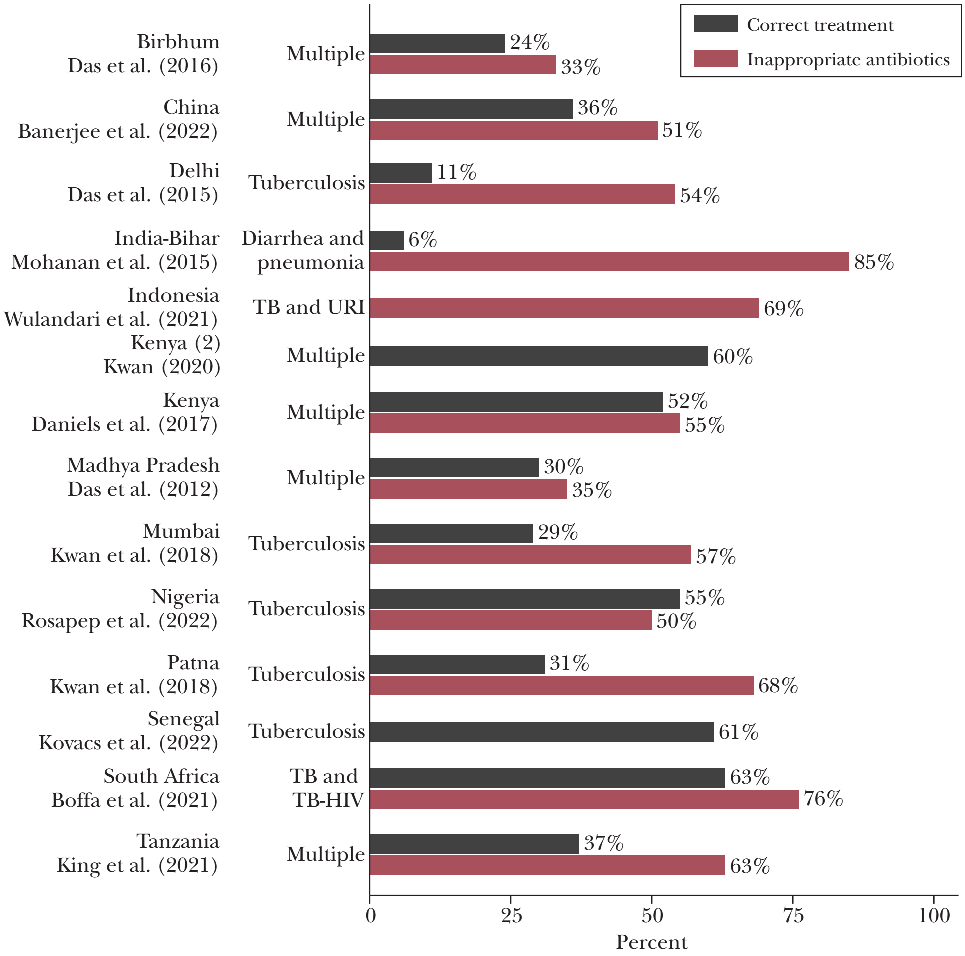
Quality of Health Care in Low- and Middle-Income Countries

The quality of health care in low- and middle-income countries can be very poor. Das et al. (2012) present one example of how standardized patients—that is, healthy people recruited from the local community and then extensively trained to present the same case to multiple providers—are treated in the Indian state of Madhya Pradesh when they present with crushing chest pain the night before and extreme anxiety.[8] After appropriate questioning, the doctor should refer this patient to a hospital, suggest an electrocardiogram, and give the patient aspirin. Das et al. (2012) show that the average interaction for this standardized patient lasts 3.5 minutes, with doctors asking three questions and completing 1.5 examinations. Das et al. (2012) suggest that a minimally correct treatment would require the doctor to complete at least one of these actions, but would not penalize the doctor for additional unnecessary tests and medicines, even if they are contraindicated or harmful. Standardized patients in their study receive such a minimally correct treatment in 31.2 percent of interactions and unnecessary or harmful treatment in 55.2 percent.

Figure 1 shows that the significant deficits in care uncovered in Das et al. (2012) hold for multiple conditions and study sites. For each study, the top bar shows the share of patients receiving the minimally correct treatment, and the bottom bar the share who received an antibiotic, which was inappropriate for all the conditions represented in the figure and is therefore a measure of an unnecessary and potentially harmful treatment. Across these studies, 40–90 percent of standardized patients are incorrectly treated, which means that they are treated for entirely the *wrong thing*—for instance, asthma or pneumonia instead of a heart attack. More stringent definitions that penalize the use of unnecessary medicines or require providers to administer *all* the required components of the treatment reduce the fraction correctly treated to less than 5 percent (and for many conditions 0–1 percent). Banerjee et al. (2023) use data from five standardized patient studies to show that one consequence of these deficits in care is that 70–85 percent of all out-of-pocket expenditures can be attributed to incorrect care or overtreatment. Interestingly, 52–78 percent of this avoidable medical expenditure is due to misdiagnosis and incorrect care rather than over treatment based on a correct diagnosis—a conclusion that holds equally for healthcare providers in the salaried public sector and in the fee-for-service private sector.

[8] Like "audit studies" in the economics of discrimination, the standardized patient approach is frequently regarded as the gold standard for measuring quality, at least for primary outpatient care. It allows researchers to abstract from omitted variable bias due to unobserved patient- and case-mix across providers and mitigates the possibility that health care providers act differently when they know they are being observed (Leonard and Masatu 2006). Most importantly, researchers can compare the physician's treatment with evidence-based clinical guidelines and evaluate the accuracy of treatment decisions even in cases where the condition is misdiagnosed, a task that is difficult to accomplish even with well-maintained patient charts as researchers do not know the true underlying illness of the patient. Studies across multiple countries and tracer conditions have shown that standardized patients can be deployed in large samples, leading to valid and reliable results with low detection rates and provider behavior consistent with the belief that they were dealing with a real patient. See Das et al. (2012) and Kwan et al. (2019).

Figure 1
Correct Management Proportions across Standardized Patient Studies



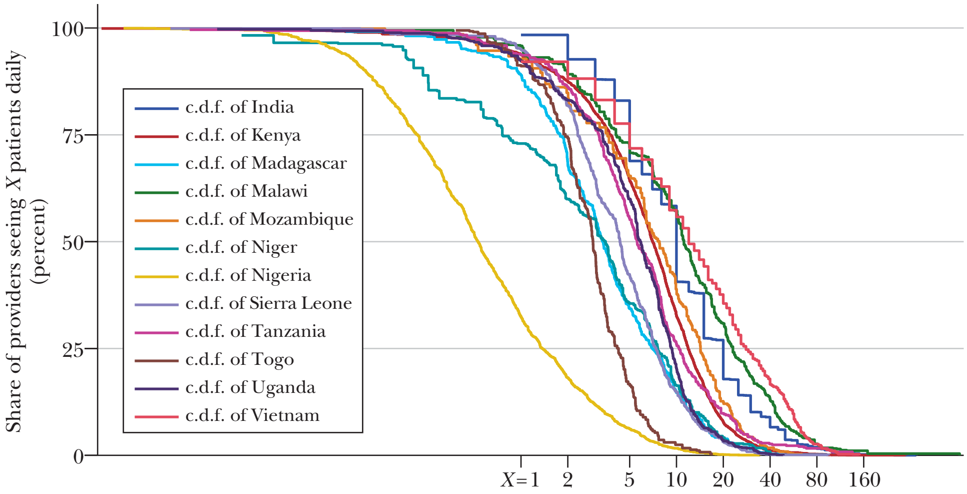
Source: Data sources listed in online Appendix C.
Notes: This figure shows the average share of standardized patients in each of 14 study sites who either (1) received at least minimal correct treatment according to study definitions; or (2) received an unnecessary antibiotic. Correct treatment was generally defined as at least one medication or test that would manage the case or advance an accurate diagnosis, regardless of whether it was completed and regardless of whether additional unnecessary or harmful tests or medications were also offered. Antibiotics are inappropriate in all cases, with the exception of diarrhea in a child (generally not measured) and a standard HRZE anti-TB antibiotic regime in a diagnosed TB case (all other antibiotics are still considered inappropriate). Bar labels show proportions of interactions with the indicated management outcome.

Capacity Constraints

Three types of capacity constraints have been evoked to explain these deficits: overcrowding, lack of equipment, and lack of adequate medical training. We consider each in turn.

Figure 2

Outpatient Capacity Utilization in 12 Low- and Middle-Income Countries



Source: The data are from Das et al. (2022) for India, World Bank and Health Strategy and Policy Institute (2016) for Vietnam, and Daniels, Das, and Gatti (2022) for the sub-Saharan African countries.
Notes: This figure shows the cumulative density functions (CDF) of health care provider daily outpatient caseloads based on facility-reported data from several studies. In each case, the per-provider-day caseload is calculated by taking the daily facility caseload and dividing by the number of providers practicing at each clinic. The CDF plots illustrate the percentage of providers in each site who are estimated to see at least the number of patients indicated on the horizontal axis each day.

Overcrowding. The World Health Organization (2016) raises frequent alarms about doctor shortages resulting in an excessive workload for health care providers in low- and middle-income countries. We are sympathetic to this explanation for certain conditions and contexts. For instance, Andrew and Vera-Hernández (2022) show that in areas with low capacity in India, demand-side incentives for women to deliver in health facilities *increased* infant mortality because the resulting congestion worsened outcomes for women with high-risk pregnancies. But taken as a whole, the utilization and capacity numbers simply do not add up to a picture of massive overcrowding.

Figure 2 shows the cumulative density functions of outpatient capacity utilization among providers in twelve low- and middle-income countries. In the country with the highest capacity utilization (Vietnam), the bottom 50 percent of providers in the patient-load distribution still see fewer than ten patients each day. In five of twelve countries, half of all providers see fewer than five patients per working day; in Nigeria, 75 percent of healthcare providers see fewer than *two* patients a day. In none of these countries do more than 25 percent of healthcare providers work an estimated full day. It would take unreasonably high estimates of the amount of time spent with each patient, or of the number of administrative and inpatient duties providers also must do, to reduce estimates of idle outpatient capacity significantly. Not surprisingly, studies that have directly examined the link between patient load and quality of care all find zero impact (Mæstad, Torsvik, and Aakvik 2010; Kovacs and Lagarde 2022; Kwan et al. 2019).[9] Indeed, given the overall data on staffing and patient loads, the real challenge here would be to explain how there could *possibly* be a causal impact of patient load when even the busiest doctors spend less than half a day (and most spend less than a couple of hours in the day) seeing patients.

Lack of equipment. A second potential explanation is the lack of infrastructure in the form of adequate facilities or medical equipment. Clearly, certain types of equipment are necessary to perform key medical functions—doctors cannot listen to a patient's heartbeat without a stethoscope. However, there have been substantial improvements in infrastructure and the availability of medical equipment in the past two decades, and it is now increasingly clear that while structural improvements are necessary for better quality care, they are far from sufficient. Multiple studies find that the correlation between the availability of medical equipment/infrastructure and quality of care is either zero or strikingly low across a range of quality measurements and in different settings. For example, Leslie, Sun, and Kruk (2017) find very low correlations between observed clinical quality with real patients and facility infrastructure for family planning, antenatal care, sick-child-care, and labor and delivery for Haiti and seven countries in sub-Saharan Africa. Similarly, Bedoya et al. (2017) report low correlations for patient safety and Das et al. (2012) for clinical quality in a standardized patient study.

Lack of medical knowledge. The third potential explanation for the poor quality of care is that healthcare providers do not have the knowledge they need to diagnose accurately and thus to treat the conditions presented to them. This explanation seems the most powerful of the three, given that studies consistently find a positive association between a doctor's quality of clinical practice and their knowledge of the case. Early contributions showed that knowing what questions to ask and examinations to perform increased the likelihood of completing these items in the clinic by 20–25 percent (Das and Hammer 2007; Das and Hammer 2014; Leonard, Masatu, and Vialou 2007; Das, Hammer, and Leonard 2008). Banerjee et al. (2023) combined tests of knowledge with standardized patients to show that knowing how to correctly manage a patient increased the likelihood of actually doing so by 22–40 percent, depending on the sample and after accounting for measurement error in measures of knowledge. At the level of associations, increasing medical knowledge improves clinical performance, but a coefficient significantly below one implies that only about one-third to one-half of improvements in clinical knowledge are then reflected in improved clinical

[9] To address the problem that demand is likely correlated with quality, Mæstad, Torsvik, and Aakvik (2010) use the size of the catchment area in Tanzania as an instrument for the clinics' caseloads, while Kovacs and Lagarde (2022) and Kwan et al. (2019) combine standardized patients with within-facility variation in caseload.

practice. Thus, medical knowledge may not be *the* binding constraint on quality of clinical care. In contrast, provider behavior conditional on medical knowledge might be this constraint—an observation we turn to next.

Provider Behavior and Market Allocation

Instead of capacity constraints in terms of patient load, poor equipment, or low clinical knowledge, two sources of inefficiency stand out as the leading explanations for poor health care quality in low- and middle-income countries. These are that doctors do not do the right thing *despite* knowing what to do and that the distribution of doctors and patterns of provider choice are such that patients do not visit the doctors who could provide them with better quality care.

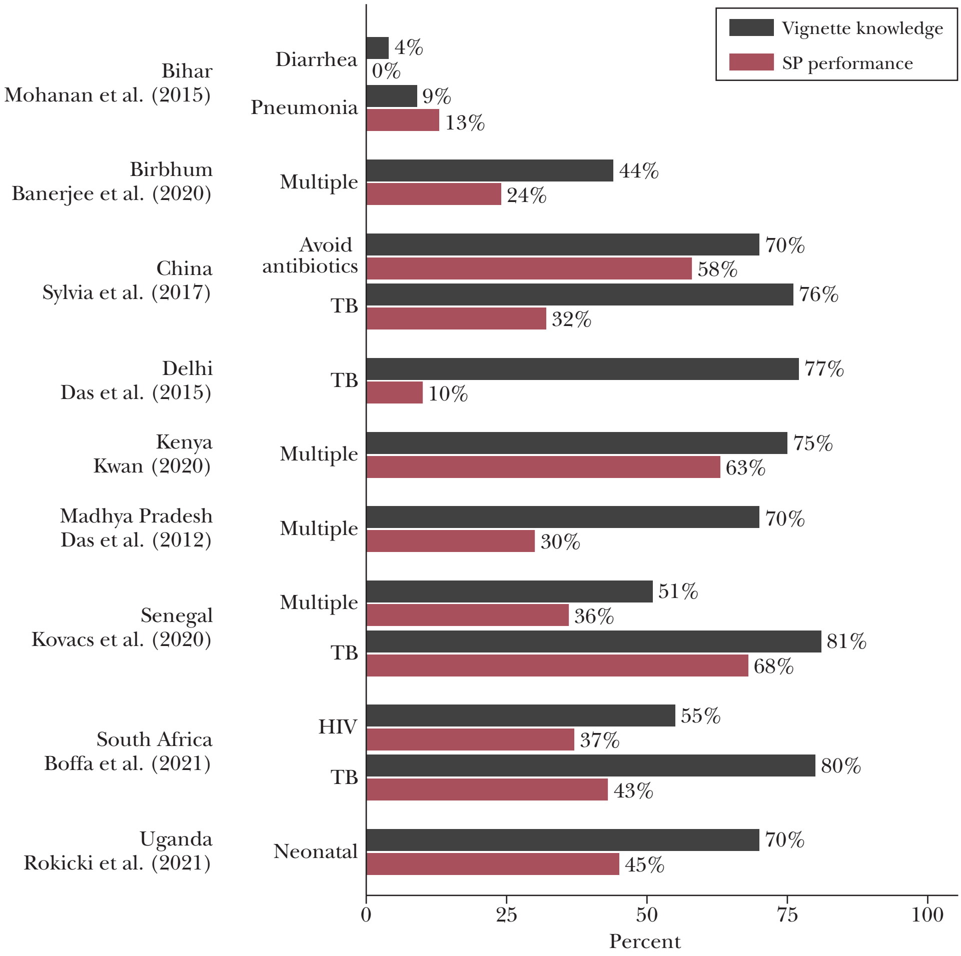
The know-do gap. Das, Hammer, and Leonard (2008) first asked whether doctors practiced at their knowledge frontier and found that they did not, a phenomenon they labelled the "know-do" gap. In Figure 3, we present the accumulated evidence from more recent studies where researchers have sent a standardized patient and later assessed the provider's knowledge of the same case. Every study (except for that of pneumonia in Bihar) finds large know-do gaps ranging from 5 to 80 percent, confirming the original findings of Das, Hammer, and Leonard (2008) across multiple countries and tracer conditions. Providers who know how to correctly treat a patient are less likely to do so in practice, and providers who (correctly) say they would avoid prescribing antibiotics (as an example of unnecessary care) are more likely to do so in practice. The know-do gap increases with medical knowledge, implying that closing the gap would offer an opportunity to significantly improve quality without investing in expensive training.

As economists would predict, the know-do gap is larger in the salaried public sector, where price incentives to provide effort are minimal. Indeed, the gaps in the public health care sector in India are so large that healthcare providers without formal training in the private sector provided similar care to fully trained doctors in the public sector, and the same doctor working in a private clinic is 42 percent more likely to treat a standardized patient correctly compared to when they are working in their public clinic (Das et al. 2016). Despite the importance of incentives, whether pay-for-performance can reduce the know-do gap remains unclear; while there were promising early results from Rwanda (Basinga et al. 2011), these have failed to replicate in a broader set of countries (for an overview of 59 studies, see Diaconu et al. 2021). In addition, research into the "know-do" gap is new, with open questions regarding the relative importance of financial incentives versus patient expectations (Currie, Lin, and Meng 2014), doctors' beliefs about patients (Banerjee et al. 2021), and behavioral biases among physicians (Groopman 2007; Mullainathan and Obermeyer 2022; Kovacs, Lagarde, and Cairns 2020) in contributing to this gap.

Patient-doctor mismatch. A second subtle source of inefficiency arises from the misallocation of patients to doctors. If there is considerable variation in doctor quality, but a low correlation between market share and quality, then the average quality received by patients will increase by inducing more visits to higher-quality

Figure 3

Know-Do Gaps between Medical Vignettes and Standardized Patients



Source: Data sources listed in online Appendix C.

Note: This figure illustrates "know-do gaps" estimated from several studies that used both medical vignettes and standardized patients with similar (or the same) samples of providers and conditions. "Vignette knowledge" is defined as the share of providers who said they would offer the patient in the indicated case scenario at least minimal correct treatment according to study definitions, regardless of what else they did (in the "Avoid antibiotics" case it is the percentage of providers who said they would not give antibiotics). "SP performance" is the proportion of providers who did the same when presented with an actual standardized patient with the same case scenario.

providers. This possibility has been explored by Daniels, Das, and Gatti (2022) for eleven sub-Saharan African countries and by Das et al. (2022) for Indian states using tests of medical knowledge (which understates the variation in clinical practice).

Two results stand out in the research on this area. First, there is indeed considerable variation in medical knowledge as measured by clinical vignettes. A general pattern is that a one standard deviation increase in a standardized index of medical knowledge increases the likelihood of knowing how to correctly treat any given vignette scenario by 10 percentage points. Thus, the difference in the general knowledge of correct treatment between a provider at the 5th percentile of the quality distribution and the 95th percentile is 40 percentage points. About 80 percent of this variation in competence is within-country or within-state. Second, the correlation between workload and clinical competence is weak. Across eleven countries in sub-Saharan Africa, against a mean of seven patients per provider per day, higher-knowledge providers have higher caseloads only in Tanzania (two additional patients per standard deviation in competence) and in Kenya and Sierra Leone (one additional patient). In Mozambique and Malawi, each additional standard deviation of provider competence is associated with two *fewer* patients per day. The remaining countries all exhibit effectively no relationship between provider knowledge and outpatient caseload.

Daniels, Das, and Gatti (2022) use data from sub-Saharan Africa and India to present a mechanical calculation of the potential gains from relocating doctors, whereby the highest quality doctors are posted to the busiest clinics, the second-highest to the second-busiest and so on, always ensuring that the relocations are within country (or state in India) and sector. They find that patient-weighted quality could increase by 0.75 standard deviations in the sub-Saharan Africa sample and by 0.5 standard deviations in the Indian sample, which corresponds to increases between 5 and 8 percentage points in the likelihood of correctly treating cases. To put the potential gains in context, they are similar to the difference in correct treatment rates for providers with and without formal training in India (3–7 percentage points) and statistically indistinguishable from a range of successful and well-powered behavioral interventions reported in Rowe et al.'s (2018) systematic review of quality improvement interventions in low- and middle-income countries.[10]

In conclusion, despite the considerable evidence that quality of health care is poor in these countries, the evidence that supply constraints are the fundamental barrier to improved health outcomes is less conclusive. Quality deficits reflect providers practicing below the knowledge frontier and allocations that result in high quality providers being underused, to the extent that they may be seeing only two to three patients a day.[11]

[10] For high-income countries, Chandra et al. (2016), show that reallocation towards higher-quality hospitals was responsible for one-third of the decline in heart attack mortality in the United States, and Dahlstrand (2022) shows that telemedicine allows doctors who are skilled at triaging to see more patients at high risk of avoidable hospitalizations in Sweden, leading to a 20 percent reduction in avoidable hospitalizations.

[11] Sparse data on hospital quality has restricted our focus to primary care, but the data that do exist suggest similar patterns. In terms of quality deficits, post-operative infections are two to three times higher in low- and middle-income countries compared to OECD countries (GlobalSurg collaborative) and, a short while after a cataract operation, patients were legally blind in 36 percent of cases (Singh, Garner, and

Why Has Health Insurance Not Improved Health Outcomes: Provider Responses?

Instead of limited capacity, we believe that the keys to understanding the uneven performance of health insurance in low- and middle-income countries is the provider side of health care.[12] Two sorts of problems arise. The first set of problems, like fraud and dispute, arise in all insurance schemes (not just health) because there are multiple transactions, each of which comes with its own potential problem. Patients must be diagnosed and treated correctly; hospitals need to submit claims and be reimbursed. Hospitals may charge false bills or overcharge for what was provided, and insurance companies may then refuse to honor the claim. In a cascading effect, insurance payment delays may lead to hospitals denying care to patients, lowering the value of insurance in the first place.

A second set of problems arises because, while physicians and hospitals may correctly charge for what they actually do, they might not perform or recommend the appropriate procedure, despite knowing what that is: the know-do gap. Health insurance alters the relative price of different procedures, potentially influencing provider behavior in a way that could result in lower health care quality.

Insurance Fraud and Administration

Health insurance fraud ranges from 3 to 15 percent of program costs in OECD countries and 3 to 10 percent in the United States (Morris 2009). It probably accounts for a larger fraction of program costs in low- and middle-income countries, as periodic reports (Ngetich 2021; Begue 2018) and audits suggest, but program-wide estimates are difficult to come by. Most countries have not made their health insurance claims data public.

Two problems that may be more salient in low- and middle-income countries are denial of care and "surprise" or double billing. Denial of care refers to a situation where insurance cards are not honored at participating hospitals. Dercon, Gunning, and Zeitlin (2019), who were the first to study this issue, show that denial of care is frequent in their setting in Kenya. Denial of care has two important implications. First, the decision to participate in the health insurance scheme depends on trust in the insurance system. Second, because denial introduces a new risk of the insurance not being honored, the overall risk with insurance may be higher than without—there is a state of the world in which the

Floyd 2000). In terms of quality variation, a study of hospital maternity wards within the single geographic area of Nairobi, Kenya, found wide variations (Siam et al. 2019). And, in terms of low capacity utilization, Colombia's National Hospitals Study in 1986 showed that the occupancy rate was 74.8 percent among Level 3 hospitals, but only 40.4 percent among Level 1 hospitals (Glassman et al. 2009).

[12] A stated objective of health insurance schemes in some countries was to allow patients to choose higher-quality providers, often in the private sector. We have documented above the existing evidence on substitution towards the private sector. However, in the light of substantial variation in quality within public and private providers, any claim on reallocation towards higher-quality providers requires facility-based measures of quality that are so far absent in the literature.

individual has paid the premium without receiving compensation when the bad state occurs. If individuals are forward-looking, this possible outcome implies that less risk-averse individuals will be *more* likely to take the insurance. By combining lab measures of trust and risk aversion with take-up decisions for an insurance product, Dercon, Gunning, and Zeitlin (2019) show that both predictions hold in their data.

Surprise or double billing is a situation whereby healthcare providers charge the insurance company the reimbursable amount, but then levy additional (and illegal) top-ups from patients. Rather than providing an administratively mandated price for a procedure, insurance reimbursements are then better regarded as a partial subsidy for the service, with pricing determined both through the usual considerations of supply and demand elasticities, but also possibly price discrimination and the special characteristics of health care markets. Again, there are no nation-level studies of double billing, because it requires surveys of insurance beneficiaries in addition to claims data. One of the few studies to combine administrative data and household surveys is Jain (2021), who studies double billing in the Indian state of Rajasthan. We will discuss this study further below.

Health Insurance and the Know-Do Gap

Do health insurance schemes affect the know-do gap? We are not aware of any studies to date in low- and middle-income countries that causally link health insurance to supply responses among providers, at least for inpatient care, where the bulk of the money is spent. It is difficult to use administrative claims data to come to any conclusion regarding quality in these settings; for discussion, see Morton et al. (2016) on how claims data are recorded. Nevertheless, we will offer an educated guess based on a collage of evidence from newspaper reports, audits, and related studies on how doctors respond to price changes in low- and middle-income countries.

As one example, media reports and field investigations from the Indian states of Andhra Pradesh, Gujarat, and Chattisgarh, shortly after the introduction of the national health insurance program RSBY, showed that many women were getting hysterectomies based on rudimentary diagnostics and for conditions such as heavy menstrual bleeding that could be medically managed. As reported by Averill and Dransfield (2013):

> A study by a non-profit organization, AP Mahila Samatha Society, in 2009 of over 1,000 women in Andhra Pradesh found an increase of 20 percent in hysterectomy cases since July 2008. They also reported that doctors had told 30 percent of the women that they would die if they did not have the operation. A few months ago, the Chhattisgarh State Health Department initiated action against 22 nursing homes, which were carrying out hysterectomies without legitimate medical reasons in order to claim money from the national health insurance scheme, Rashtriya Swasthya Bima Yojana (RSBY).

Subsequent research indeed confirms much higher rates of hysterectomies in the states of Gujarat, Bihar, and Andhra Pradesh, but also cautions that causal claims on the impact of insurance are harder to establish (Desai, Sinha, and Mahal 2011; Desai et al. 2019).

Cataract surgery seems to be another area with sharp increases after the arrival of health insurance. For instance, Rana (2017) reports that in the Indian state of West Bengal:

> Private facilities were found to concentrate mainly on easy-to-handle services, like cataract surgery, and commission agents recruited patients for this surgery, often without indication that the patient even needed the surgery. From the 1,090 procedures performed under RSBY, I found that the actual treatment done by the private hospitals occurred not to provide health care for patients, but instead to profit for health care facilities. It also involved a huge informational asymmetry, as it seemed to be impossible for the patient to keep track of as to which of the 1,090 procedures covered by RSBY was performed on him or her.

Jain (2021) is the first study from a low- or middle-income country that looks at hospital pricing and coding systematically in the context of an insurance scheme. She combines administrative claims data with a large household survey for the Indian state of Rajasthan, which allows her to better understand how hospitals react to changes in administrative prices. Without the household survey, for instance, it would have been impossible to determine how much households are asked to pay out-of-pocket because the practice is illegal and therefore off-the-books. She finds that providers do not respect administrative prices: 41 percent of patients paid for their treatment even though the care was supposed to be free and the average payments were $35, which is a large sum for poor households and represents a 37 percent increase over the insurance reimbursement rate.

Moreover, hospitals react rapidly to adjustments in reimbursement rates. Jain (2021) finds that with every additional ₹100 in reimbursements, prices charged to patients decreased—but only by ₹55. She also uses an event-study to show that when the relative reimbursement rates within a category change (for instance, childbirth with and without an episiotomy), so do the reported procedures. Within a week of a price change, a 1 percent increase in the reimbursement rate induced a 0.4 percent increase in its claim volume. She suggests that this reflects “up-coding,” whereby health care providers submit codes for more expensive care than actually provided, but a bigger worry, which she does not rule out, is that hospitals changed the treatments that patients received.

Other studies provide systematic evidence on differences in quality of care by insurance status, at least for outpatient care. One set of studies finds that when patients get health insurance, their satisfaction remains the same or worsens (Bauhoff, Hotchkiss, and Smith 2011; Robyn et al. 2013). Two studies have used standardized patients, varying their insurance status across visits. In South Africa, Sripathy (2020) shows that clinical effort is higher for standardized patients with insurance, but the proportion correctly treated does not change and the extent of unnecessary and more expensive treatments increases. In China, Lu (2014) finds a similar result, with the greater use of unnecessary antibiotics for insured patients, but only when doctors have a direct financial incentive associated with the purchase of the medicine. In the Philippines, Gertler and Solon (2000) find that because hospitals do not charge the mandated administrative price, insured patients pay 86 percent of what the uninsured would pay, sharply limiting any financial protection offered by the scheme. Finally, in Burkina Faso, Fink et al. (2013) find that in the context of a capitation-based payment system, quality of care was significantly lower for insured patients in participating health facilities compared to those who were not insured.

These studies are not cherry-picked; they constitute the full corpus of what we have found in the literature on provider behavior in response to insurance in low- and middle-income countries. The existence of so few studies on this key subject is itself a cause for concern. The fact that every study showed that provider behavior undermined the objectives of the scheme and contributed to an increase in the know-do gap is an even bigger worry.

Discussion and Conclusion

A considerable literature from low- and middle-income countries over the last two decades highlights several noteworthy features of health insurance schemes. In terms of the structure, governments have converged on using public subsidies for health insurance premiums, which are now nominally priced or free in most countries. On the other hand, governments have diverged in how they reimburse providers for services, using a wide range of payment mechanisms that are frequently revised and overhauled. In terms of outcomes, the schemes have provided financial protection with a decline in out-of-pocket expenditures, but these gains have not translated into demand for unsubsidized health insurance. Furthermore, these schemes tend to increase utilization without a concomitant improvement in health outcomes. Finally, the lack of consistent improvements in outcomes is not because of supply constraints in terms of workload, equipment, or knowledge, but instead due to behavioral responses on the part of providers. Health insurance does not systematically improve the quality of existing providers, and often seems to make it worse. There is also little evidence to show that health insurance allows patients to visit higher-quality providers.

The phenomena of low demand, poor health outcomes and adverse behavioral responses, while seemingly disparate, are consistent with an underlying framework that recognizes the special features of healthcare. While adverse selection is traditionally regarded as the defining unique feature of health insurance, once insurance premiums are tax-funded, it is less of a concern. Instead, what is germane here is the "credence good" aspect of healthcare, whereby physicians know what patients need, but patients (and health insurance companies) do not. This informational asymmetry leads to overtreatment if patients are treated for serious problems when their condition is mild, and undertreatment or incorrect treatment if patients are treated for a mild condition when their condition is serious. Both are inefficient, as insurance pays for unnecessary treatment in the case of overtreatment, and patients lose the surplus from good health in the case of undertreatment. Because physicians enjoy considerable latitude in choosing the treatment, they may distort treatment decisions in a manner that is beneficial to themselves rather than to the patient.

Theoretically, the dual inefficiencies of over- and undertreatment can be alleviated through a combination of price and nonprice incentives. The latter include enhancing altruistic motives, professionalism, peer reviews, and a host of norms and principles. Interestingly, even in the absence of nonprice mechanisms, price incentives alone can deliver efficient outcomes in markets with credence goods under certain conditions (Dulleck and Kerschbamer 2006).[13] In practice however, accurate price-setting requires a high degree of transaction and physician-specific information, which is unlikely to be available for administrators in any insurance scheme. Consequently, we see countries adjusting their pricing mechanisms as providers exploit deficiencies in existing purchasing agreements; we see little improvement in health outcomes despite increased utilization because of increasing unnecessary care (cataracts, hysterectomies) and a possible decline in the quality of each interaction; and we see systematic changes in provider behavior that undermine the stated objectives of the insurance scheme.

This idea—that health insurance affects both demand for and supply of quality health care—is not new; for example, Arrow's article on healthcare six decades ago dealt with the doctor-patient relationship and the problem of trust or credence (Arrow 1963). More recently, Newhouse (2014) considers the role of provider moral hazard in explaining why the US-based Rand Health Insurance Experiment showed that more insurance led to increased utilization, but not improved health outcomes—a result similar to what we have documented here. Newhouse wrote: "[T]he odds that a service at the margin helped them were probably offset by the odds that it hurt them. I have felt more confidence in this explanation over time as evidence of medical error and poor quality of care has piled up . . ." Indeed, our review uncovered multiple papers that sought to explain why insurance does not improve health outcomes by pointing to the poor administration of the scheme or unexpected departures from what the scheme was supposed to do.

Moving forward, in terms of the research, future studies using demand-side data can still be insightful for several open questions. Does financial protection alone provide sufficient justification for expanding health insurance (Finkelstein,

[13] What disciplines doctors in this case is physician-specific pricing that equalizes the markups from different treatments. This is because posted prices reveal information about the doctors' strategy: if costs are known, patients correctly infer that a physician will always choose the treatment plan that offers a higher profit. This predictability in turn implies that there is no further information asymmetry and therefore no incentive for the doctor to distort her behavior in order to extract surplus from the patient.

Hendren, and Luttmer 2019)? Does low demand reflect an actuarial calculation or administrative burdens or other costs, perhaps linked to behavioral issues? Does health insurance lead to improved health outcomes in studies with sufficiently large sample sizes and a broad set of indicators?

But while these questions are important, where we desperately need new evidence is instead on the supply side of the market where the major failures are concentrated. If our diagnosis of the problems of health insurance in low- and middle-income countries is correct, the key questions are (1) whether the arrival of health insurance allows households to visit higher-quality facilities and (2) whether the arrival of health insurance increases the quality of clinical interactions among existing providers. We have not found any studies that causally link health insurance to objectively-measured higher-quality choices (as opposed to proxy measures, such as using "private" or "public" as measures of quality) or document supply responses to the arrival of health insurance. Providing this evidence is admittedly not easy; for example, data on post-hospitalization outcomes requires teams to track hospital users to their homes months after their procedure. Yet these two questions are where we will likely see the largest gains in our understanding of how (and whether) health insurance can improve the health of populations in low- and middle-income countries.

We cannot separate health insurance from the quality of care, nor can we separate quality of care from specific reimbursement mechanisms. Consequently, the issue at heart is *not* whether government subsidies should be channeled through health insurance premiums or direct subsidies to public facilities. Instead, the question is what specific payment structures and nonprice mechanisms can alter provider behavior and patient choice to improve quality under any administrative regime.

■ *We dedicate this paper to the memory of our colleague and friend, Adam Wagstaff, whose work on health insurance presaged many of the developments presented here. We wish he had been around to read and argue about this article. We thank Benjamin Daniels and Samikshya Siwakoti for exceptional research assistance. We thank Jessica Cohen, Stefan Dercon, Shanta Devarajan, Günther Fink, Anne Fitzpatrick, Kara Hanson, Radhika Jain, Eeshani Kandpal, Margaret McConnell, Anne Mills, Grant Miller, Aakash Mohpal, Manoj Mohanan, Timothy Powell-Jackson, Edward Okeke, Rebecca Thornton, Marcos Vera-Hernandez, and Andrew Zeitlin for early inputs. We especially thank Jeffrey Hammer for detailed comments on an earlier draft. Timothy Taylor and the editors of this journal were instrumental in helping us refine and tighten our arguments. We acknowledge funding from the World Bank's Research Support Budget, for project P178514. The findings, interpretations, and conclusions expressed here are those of the authors and do not necessarily represent the views of the World Bank, its executive directors, or the governments they represent.*

References

Amu, Hubert, Kwamena Sekyi Dickson, Kenneth Setorwu Adde, Kwaku Kissah-Korsah, Eugene Kofuor Maafo Darteh, and Akwasi Kumi-Kyereme. 2022. "Prevalence and Factors Associated with Health Insurance Coverage in Urban Sub-Saharan Africa: Multilevel Analyses of Demographic and Health Survey Data." *Plos One* 17 (3): e0264162.

Andrew, Alison, and Marcos Vera-Hernández. 2022. "Incentivizing Demand for Supply-Constrained Care: Institutional Birth in India." *Review of Economics and Statistics*. https://doi.org/10.1162/rest_a_01206.

Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* 53 (5): 941–73.

Averill, Ceri, and Sarah Dransfield. 2013. "Unregulated and Unaccountable: How the Private Health Care Sector in India Is Putting Women's Lives at Risk." *Oxfam International*, February. https://policy-practice.oxfam.org/resources/unregulated-and-unaccountable-how-the-private-health-care-sector-in-india-is-pu-268392/.

Barber, Sarah L., and Luca Lorenzoni, and Paul Ong. 2019. *Price Setting and Price Regulation in Health Care: Lessons for Advancing Universal Health Coverage*. World Health Organization. https://apps.who.int/iris/handle/10665/325547.

Baker, Laurence C. 2010. "Acquisition of MRI Equipment by Doctors Drives Up Imaging Use and Spending." *Health Affairs* 29 (12): 2252–59.

Balsa, Ana I., and Patricia Triunfo. 2021. "The Effects of Expanded Social Health Insurance on Young Mothers: Lessons from a Pro-choice Reform in Uruguay." *Health Economics* 30 (3): 603–22.

Banerjee, Abhijit, Esther Duflo, and Richard Hornbeck. 2014. "Bundling Health Insurance and Microfinance in India: There Cannot Be Adverse Selection If There Is No Demand." *American Economic Review* 104 (5): 291–97.

Banerjee, Abhijit, Jishnu Das, Jeffrey Hammer, Reshmaan Hussam, and Aakash Mohpal. 2023. "The Market for Healthcare in Low Income Countries." Unpublished.

Banerjee, Abhijit, Amy Finkelstein, Rema Hanna, Benjamin A. Olken, Arianna Ornaghi, and Sudarno Sumarto. 2021. "The Challenges of Universal Health Insurance in Developing Countries: Experimental Evidence from Indonesia's National Health Insurance." *American Economic Review* 111 (9): 3035–63.

Barasa, Edwine, Jacob Kazungu, Peter Nguhiu, and Nirmala Ravishankar. 2021. "Examining the Level and Inequality in Health Insurance Coverage in 36 Sub-Saharan African Countries." *BMJ Global Health* 6 (4): e004712.

Barasa, Edwine, Khama Rogo, Njeri Mwaura, and Jane Chuma. 2018. "Kenya National Hospital Insurance Fund Reforms: Implications and Lessons for Universal Health Coverage." *Health Systems and Reform* 4 (4): 346–61.

Basinga, Paulin, Paul J. Gertler, Agnes Binagwaho, Agnes L. B. Soucat, Jennifer Sturdy, and Christel M. J. Vermeersch. 2011. "Effect on Maternal and Child Health Services in Rwanda of Payment to Primary Health-Care Providers for Performance: An Impact Evaluation." *The Lancet* 377 (9775): 1421–28.

Bauhoff, Sebastian, David R. Hotchkiss, and Owen Smith. 2011. "The Impact of Medical Insurance for the Poor in Georgia: A Regression Discontinuity Approach." *Health Economics* 20 (11): 1362–78.

Bedoya, Guadalupe, Amy Dolinger, Khama Rogo, Njeri Mwaura, Francis Wafula, Jorge Coarasa, Ana Goicoechea, and Jishnu Das. 2017. "Observations of Infection Prevention and Control Practices in Primary Health Care, Kenya." *Bulletin of the World Health Organization* 95 (7): 503–16.

Begue, Michelle. 2018. "Colombia Battles Corruption within Healthcare System." *CGTN America*, November 20. https://america.cgtn.com/2018/11/20/colombia-battles-corruption-within-healthcare-system.

Bernal, Noelia, Miguel A. Carpio, and Tobias J. Klein. 2017. "The Effects of Access to Health Insurance: Evidence from a Regression Discontinuity Design in Peru." *Journal of Public Economics* 154: 122–36.

Blanchet, Nathan J., Günther Fink, and Isaac Osei-Akoto. 2012. "The Effect of Ghana's National Health Insurance Scheme on Health Care Utilisation." *Ghana Medical Journal* 46 (2): 76–84.

Camacho, Adriana, and Emily Conover. 2013. "Effects of Subsidized Health Insurance on Newborn Health in a Developing Country." *Economic Development and Cultural Change* 61 (3): 633–58.

Capuno, Joseph J., Aleli D. Kraft, Stella Quimbo, Carlos R. Tan Jr., and Adam Wagstaff. 2016. "Effects of

Price, Information, and Transactions Cost Interventions to Raise Voluntary Enrollment in a Social Health Insurance Scheme: A Randomized Experiment in the Philippines." *Health Economics* 25 (6): 650–62.

Celhay, Pablo A., Paul J. Gertler, Paula Giovagnoli, and Christel Vermeersch. 2019. "Long-Run Effects of Temporary Incentives on Medical Care Productivity." *American Economic Journal: Applied Economics* 11 (3): 92–127.

Chandra, Amitabh, Amy Finkelstein, Adam Sacarny, and Chad Syverson. 2016. "Health Care Exceptionalism? Performance and Allocation in the US Health Care Sector." *American Economic Review* 106 (8): 2110–44.

Clemens, Jeffrey, and Joshua D. Gottlieb. 2014. "Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?" *American Economic Review* 104 (4): 1320–49.

Cohen, François, and Antoine Dechezleprêtre. 2022. "Mortality, Temperature, and Public Health Provision: Evidence from Mexico." *American Economic Journal: Economic Policy* 14 (2): 161-92.

Currie, Janet, Wanchuan Lin, and Juanjuan Meng. 2014. "Addressing Antibiotic Abuse in China: An Experimental Audit Study." *Journal of Development Economics* 110: 39–51.

Dahlstrand, Amanda. 2022. "Defying Distance? The Provision of Services in the Digital Age." https://www.dropbox.com/s/skusfyobfaoku2c/Dahlstrand_JMP.pdf?dl=0.

Daniels, Benjamin, Jishnu Das, and Roberta Gatti. 2022. "Analysis of SDI Health Data (Version V0)." https://doi.org/10.5281/zenodo.6478472 (accessed February 17, 2023).

Das, Jishnu, and Quy-Toan Do. 2023. "Replication data for: The Prices in the Crises: What We Are Learning from 20 Years of Health Insurance in Low- and Middle-Income Countries." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E184523V1.

Das, Jishnu, and Jeffrey Hammer. 2007. "Money for Nothing: The Dire Straits of Medical Practice in Delhi, India." *Journal of Development Economics* 83 (1): 1–36.

Das, Jishnu, and Jeffrey Hammer. 2014. "Quality of Primary Care in Low-Income Countries: Facts and Economics." *Annual Review of Economics* 6: 525–53.

Das, Jishnu, Jeffrey Hammer, and Kenneth Leonard. 2008. "The Quality of Medical Advice in Low-Income Countries." *Journal of Economic Perspectives* 22 (2): 93–114.

Das, Jishnu, Alaka Holla, Aakash Mohpal, and Karthik Muralidharan. 2016. "Quality and Accountability in Health Care Delivery: Audit-Study Evidence from Primary Care in India." *American Economic Review* 106 (12): 3765–99.

Das, Jishnu, Benjamin Daniels, Monisha Ashok, Eun-Young Shim, and Karthik Muralidharan. 2022. "Two Indias: The Structure of Primary Health Care Markets in Rural Indian Villages with Implications for Policy." *Social Science and Medicine* 112799.

Das, Jishnu, Alaka Holla, Veena Das, Manoj Mohanan, Diana Tabak, and Brian Chan. 2012. "In Urban and Rural India, a Standardized Patient Study Showed Low Levels of Provider Training and Huge Quality Gaps." *Health Affairs* 31 (12): 2774–84.

Dercon, Stefan, Jan Willem Gunning, and Andrew Zeitlin. 2019. "The Demand for Insurance under Limited Trust: Evidence from a Field Experiment in Kenya." Centre for the Study of African Economies, University of Oxford Working Paper 2019-06.

Desai, Sapna, Ankita Shukla, Devaki Nambiar, and Rajani Ved. 2019. "Patterns of Hysterectomy in India: A National and State-Level Analysis of the Fourth National Family Health Survey (2015–2016)." *BJOG: An International Journal of Obstetrics and Gynaecology* 126 (S4): 72–80.

Desai, Sapna, Tara Sinha, and Ajay Mahal. 2011. "Prevalence of Hysterectomy among Rural and Urban Women with and without Health Insurance in Gujarat, India." *Reproductive Health Matters* 19 (37): 42–51.

DHS Program. "Demographic and Health Surveys." Various years. USAID. Rockville: Inner City Fund.

Diaconu, Karin, Jennifer Falconer, Adrian Verbel, Atle Fretheim, and Sophie Witter. 2021. "Paying for Performance to Improve the Delivery of Health Interventions in Low- and Middle-Income Countries." *Cochrane Database of Systematic Reviews* 5: CD007899.

Drechsler, Denis, and Johannes Jütting. 2005. "Private Health Insurance for the Poor in Developing Countries?" *OECD Development Centre Policy Insights* 11.

Dulleck, Uwe, and Rudolf Kerschbamer. 2006. "On Doctors, Mechanics, and Computer Specialists: The Economics of Credence Goods." *Journal of Economic Literature* 44 (1): 5–42.

Escobar, María-Luisa, Ursula Giedion, Antonio Giuffrida, and Amanda L. Glassman. 2010. "Colombia: After a Decade of Health System Reform." In *From Few to Many: Ten Years of Health Insurance*

Expansion in Colombia, 1–13. New York: Inter-American Development Bank.

Fink, Günther, Paul Jacob Robyn, Ali Sié, and Rainer Sauerborn. 2013. "Does Health Insurance Improve Health? Evidence from a Randomized Community-Based Insurance Rollout in Rural Burkina Faso." *Journal of Health Economics* 32 (6): 1043–56.

Finkelstein, Amy, Nathaniel Hendren, and Erzo F. P. Luttmer. 2019. "The Value of Medicaid: Interpreting Results from the Oregon Health Insurance Experiment." *Journal of Political Economy* 127 (6): 2836–74.

Fitzpatrick, Anne, and Rebecca Thornton. 2019. "The Effects of Health Insurance within Families: Experimental Evidence from Nicaragua." *World Bank Economic Review* 33 (3): 736–49.

Garg, Samir, Kirtti Kumar Bebarta, and Narayan Tripathi. 2020. "Performance of India's National Publicly Funded Health Insurance Scheme, Pradhan Mantri Jan Arogaya Yojana (PMJAY), in Improving Access and Financial Protection for Hospital Care: Findings from Household Surveys in Chhattisgarh State." *BMC Public Health* 20: 949.

Garg, Samir, Sayantan Chowdhury, and T. Sundararaman. 2019. "Utilisation and Financial Protection for Hospital Care under Publicly Funded Health Insurance in Three States in Southern India." *BMC Health Services Research* 19: 1004.

Gertler, Paul, and Jonathan Gruber. 2002. "Insuring Consumption against Illness." *American Economic Review* 92 (1): 51–70.

Gertler, Paul, and Orville Solon. 2000. "Who Benefits from Social Health Insurance in Developing Countries." Unpublished.

Giedion, Ursula, and Manuela Villar Uribe. 2009. "Colombia's Universal Health Insurance System." *Health Affairs* 28 (3): 853–63.

Glassman, Amanda L., María-Luisa Escobar, Antonio Giuffrida, Ursula Giedion, eds. 2009. *From Few to Many: Ten Years of Health Insurance Expansion in Colombia.* Washington, DC: Inter-American Development Bank.

Grépin, Karen A. 2016. "Private Sector an Important but Not Dominant Provider of Key Health Services in Low- And Middle-Income Countries." *Health Affairs* 35 (7): 1214–21.

Groopman, Jerome E. 2007. *How Doctors Think.* Boston: Houghton Mifflin.

Gruber, Jonathan, and Maria Owings. 1996. "Physician Financial Incentives and Cesarean Section Delivery." *RAND Journal of Economics* 27 (1): 99–123.

Hanson, Kara, Edwine Barasa, Ayako Honda, Warisa Panichkriangkrai, and Walaiporn Patcharanarumol. 2019. "Strategic Purchasing: The Neglected Health Financing Function for Pursuing Universal Health Coverage in Low- and Middle-Income Countries." *International Journal of Health Policy and Management* 8 (8): 501–04.

Inter-American Development Bank. 2015. "Strengthening Governments Capacity to Discern Value the Need to Address Technological Pressure on Health Expenditure." Inter-American Development Bank. https://publications.iadb.org/publications/english/document/Breve-13-Strengthening-Governments-Capacity-to-Discern-Value-the-Need-to-Address-Technological-Pressure-on-Health-Expenditure.pdf.

Jain, Radhika. 2021. "Private Hospital Behavior under Government Health Insurance: Evidence from India." Paper presented at iHEA World Congress on Health Economics, July 13.

Karan, Anup, Winnie Yip, and Ajay Mahal. 2017. "Extending Health Insurance to the Poor in India: An Impact Evaluation of Rashtriya Swasthya Bima Yojana on Out-of-Pocket Spending for Healthcare." *Social Science and Medicine* 181: 83–92.

King, Gary, Emmanuela Gakidou, Kosuke Imai, Jason Lakin, Ryan T. Moore, Clayton Nall, Nirmala Ravishankar, et al. 2009. "Public Policy for the Poor? A Randomised Assessment of the Mexican Universal Health Insurance Programme." *The Lancet* 373 (9673): 1447–54.

Kovacs, Roxanne, and Mylene Lagarde. 2022. "Does High Workload Reduce the Quality of Healthcare? Evidence from Rural Senegal." *Journal of Health Economics* 82: 102600.

Kovacs, Roxanne J., Mylene Lagarde, and John Cairns. 2020. "Overconfident Health Workers Provide Lower Quality Healthcare." *Journal of Economic Psychology* 76: 102213.

Kwan, Ada, Benjamin Daniels, Sofi Bergkvist, Veena Das, Madhukar Pai, and Jishnu Das. 2019. "Use of Standardised Patients for Healthcare Quality Research in Low- and Middle-Income Countries." *BMJ Global Health* 4 (5): e001669.

Leonard, Kenneth, and Melkiory C. Masatu. 2006. "Outpatient Process Quality Evaluation and the Hawthorne Effect." *Social Science and Medicine* 63 (9): 2330–40.

Leonard, Kenneth L., Melkiory C. Masatu, and Alexandre Vialou. 2007. "Getting Doctors to Do Their

Best: The Roles of Ability and Motivation in Health Care Quality." *Journal of Human Resources* 42 (3): 682–700.

Levine, David, Rachel Polimeni, and Ian Ramage. 2016. "Insuring Health or Insuring Wealth? An Experimental Evaluation of Health Insurance in Rural Cambodia." *Journal of Development Economics* 119: 1–15.

Leslie, Hannah H., Zeye Sun, and Margaret E. Kruk. 2017. "Association between Infrastructure and Observed Quality of Care in 4 Healthcare Services: A Cross-Sectional Study of 4,300 Facilities in 8 Countries." *PLoS Medicine* 14 (12): e1002464.

Liang, Xiaoyun, Hong Guo, Chenggang Jin, Xiaoxia Peng, and Xiulan Zhang. 2012. "The Effect of New Cooperative Medical Scheme on Health Outcomes and Alleviating Catastrophic Health Expenditure in China: A Systematic Review." *PloS One* 7 (8): e40850.

Londoño, Juan-Luis, and Julio Frenk. 1997. "Structured Pluralism: Towards an Innovative Model for Health System Reform in Latin America." *Health Policy* 41 (1): 1–36.

Lu, Fangwen. 2014. "Insurance Coverage and Agency Problems in Doctor Prescriptions: Evidence from a Field Experiment in China." *Journal of Development Economics* 106: 156–67.

Mæstad, Ottar, Gaute Torsvik, and Arild Aakvik. 2010. "Overworked? On the Relationship between Workload and Health Worker Performance." *Journal of Health Economics* 29 (5): 686–98.

Malani, Anup, Phoebe Holtzman, Kosuke Imai, Cynthia Kinnan, Morgen Miller, Shailender Swaminathan, Alessandra Voena, Bartosz Woda, and Gabriella Conti. 2021. "Effect of Health Insurance in India: A Randomized Controlled Trial." NBER Working Paper 29576.

Miller, Grant, Diana Pinto, and Marcos Vera-Hernández. 2013. "Risk Protection, Service Use, and Health Outcomes under Colombia's Health Insurance Program for the Poor." *American Economic Journal: Applied Economics* 5 (4): 61–91.

Morris, Lewis. 2009. "Combating Fraud in Health Care: An Essential Component of Any Cost Containment Strategy." *Health Affairs* 28 (5): 1351–56.

Morton, Matthew, Somil Nagpal, Rajeev Sadanandan, and Sebastian Bauhoff. 2016. "India's Largest Hospital Insurance Program Faces Challenges in Using Claims Data to Measure Quality." *Health Affairs* 35 (10): 1792–99.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *Quarterly Journal of Economics* 137 (2): 679–727.

National Population Commission and ICF International. 2014. "Nigeria Demographic and Health Survey 2013." Abuja, Nigeria and Rockville, MD: NPC Nigeria and ICF International.

Newhouse, Joseph P. 2014. "Introduction." In *Moral Hazard in Health Insurance*, 1–12. New York: Columbia University Press.

Ngetich, Jacob. 2021. "NHIF Loses Sh10 Billion through Fake Claims from Health Facilities." *The Standard* (*Kenya*), September 13. https://www.standardmedia.co.ke/national/article/2001423351/nhif-loses-sh10-billion-through-fake-claims-from-health-facilities.

Ortiz-Ospina, Esteban, and Max Roser. 2017. "Healthcare Spending." *Our World in Data*, June. https://ourworldindata.org/financing-healthcare.

Palacios, Robert, Jishnu Das, and Changqing Sun. 2011. *India's Health Insurance Scheme for the Poor. Evidence from the Early Experience of the Rashtriya Swasthya Bima Yojana.* New Delhi: Centre for Policy Research.

Pauly, Mark V., Peter Zweifel, Richard M. Scheffler, Alexander S. Preker, and Mark Bassett. 2006. "Private Health Insurance in Developing Countries." *Health Affairs* 25 (2): 369–79.

Pettigrew, Luisa M., and Inke Mathauer. 2016. "Voluntary Health Insurance Expenditure in Low- and Middle-Income Countries: Exploring Trends during 1995–2012 and Policy Implications for Progress towards Universal Health Coverage." *International Journal for Equity in Health* 15: 67.

Powell-Jackson, Timothy, Kara Hanson, Christopher J. M. Whitty, and Evelyn K. Ansah. 2014. "Who Benefits from Free Healthcare? Evidence from a Randomized Experiment in Ghana." *Journal of Development Economics* 107: 305–19.

Rana, Kumar. 2017. "Health Insurance for the Poor, or Privatization by Stealth? A Study on the Rashtriya Swasthya Bima Yojana (RSBY) in India." PhD diss. Harvard University.

Raza, Wameq A., Ellen van de Poel, Arjun Bedi, and Frans Rutten. 2016. "Impact of Community-Based Health Insurance on Access and Financial Protection: Evidence from Three Randomized Control Trials in Rural India." *Health Economics* 25 (6): 675–87.

Robyn, Paul Jacob, Till Bärnighausen, Aurélia Souares, Germain Savadogo, Brice Bicaba, Ali Sié, and Rainer Sauerborn. 2013. "Does Enrollment Status in Community-Based Insurance Lead to Poorer Quality of Care? Evidence from Burkina Faso." *International Journal for Equity in Health* 12: 31.

Romero, Mauricio. 2014. "Financial Incentives and Pharmaceutical Prices: The Colombian Case." Unpublished.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2022. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." arXiv: 2201.01194.

Rowe, Alexander K., Samantha Y. Rowe, David H. Peters, Kathleen A. Holloway, John Chalker, and Dennis Ross-Degnan. 2018. "Effectiveness of Strategies to Improve Health-Care Provider Practices in Low-Income and Middle-Income Countries: A Systematic Review." *The Lancet Global Health* 6 (11): e1163–75.

Scheil-Adlung, Xenia. 2014. "Universal Health Protection: Progress to Date and the Way Forward." International Labour Organization Social Protection Policy Paper 10.

Siam, Zeina Ali, Margaret McConnell, Ginger Golub, George Nyakora, Claire Rothschild, and Jessica Cohen. 2019. "Accuracy of Patient Perceptions of Maternity Facility Quality and the Choice of Providers in Nairobi, Kenya: A Cohort Study." *BMJ Open* 9 (7): e029486.

Singh, A. J., Paul Garner, and Katherine Floyd. 2000. "Cost-Effectiveness of Public-Funded Options for Cataract Surgery in Mysore, India." *The Lancet* 355 (9199): 180–84.

Somanathan, Aparnaa, Ajay Tandon, Huong Lan Dao, Kari L. Hurt, and Hernan L. Fuenzalida-Puelma. 2014. *Moving toward Universal Coverage of Social Health Insurance in Vietnam: Assessment and Options.* Washington, DC: World Bank.

Sood, Neeraj, and Zachary Wagner. 2018. "Social Health Insurance for the Poor: Lessons from a Health Insurance Programme in Karnataka, India." *BMJ Global Health* 3 (1): e000582.

Spence, Michael, and Richard Zeckhauser. 1971. "Insurance, Information, and Individual Action." *American Economic Review* 61 (2): 380–87.

Spenkuch, Jörg L. 2012. "Moral Hazard and Selection among the Poor: Evidence from a Randomized Experiment." *Journal of Health Economics* 31 (1): 72–85.

Sripathy, Arthika. 2020. "Rationalising Health Care Provision under Market Incentives: Experimental Evidence from South Africa." PhD diss. London School of Economics and Political Science.

Tanzi, Vito, and Ludger Schuknecht. 2000. *Public Spending in the 20th Century: A Global Perspective.* Cambridge, UK: Cambridge University Press.

Thornton, Rebecca L. 2021. "Boosting Enrollment in Health Insurance through Subsidies and Enrollment Assistance: What Works, and Who Is Selected?" Unpublished.

Thornton, Rebecca L., Laurel E. Hatt, Erica M. Field, Mursaleena Islam, Freddy Solís Diaz, and Martha Azucena González. 2010. "Social Security Health Insurance for the Informal Sector in Nicaragua: A Randomized Evaluation." *Health Economics* 19 (S): 181–206.

Tien, Tran Van, Hoang Thi Phuong, Inke Mathauer, and Nguyen Thi Kim Phuong. 2011. *A Health Financing Review of Viet Nam with a Focus on Social Health Insurance: Bottlenecks in Institutional Design and Organizational Practice of Health Financing and Options to Accelerate Progress towards Universal Coverage.* World Health Organization. https://apps.who.int/iris/handle/10665/341160.

Tussing, A. Dale, Hongying Wang, and Yidong Wang. 2014. "Chinese Doctors in Crisis: Discontented and in Danger." *Health Affairs*, May 27. https://www.healthaffairs.org/do/10.1377/forefront.20140527.039172/full/.

Wagstaff, Adam, Ha Thi Hong Nguyen, Huyen Dao, and Sarah Bales. 2016. "Encouraging Health Insurance for the Informal Sector: A Cluster Randomized Experiment in Vietnam." *Health Economics* 25 (6): 663–74.

Wang, Wenjuan, Gheda Temsah, and Lindsay Mallick. 2014. *Health Insurance Coverage and Its Impact on Maternal Health Care Utilization in Low and Middle-Income Countries.* Rockville, MD: ICF International.

World Bank, and Health Strategy and Policy Institute. 2016. "Quality and Equity in Basic Health Care Services in Vietnam: Findings from the Vietnam District and Commune Health Facility Survey." Washington, DC: World Bank.

World Health Organization. 2000. *The World Health Report 2000: Health Systems—Improving Performance.* World Health Organization. https://apps.who.int/iris/handle/10665/42281.

World Health Organization. 2016. "Health Workforce Requirements for Universal Health Coverage and the Sustainable Development Goals." World Health Organization Human Resources for Health Observer 17.

World Health Organization. 2020. *Private Sector Landscape in Mixed Health Systems.* World Health Organization. https://apps.who.int/iris/handle/10665/353170.

America's Continuing Struggle with Mental Illnesses: Economic Considerations

Richard G. Frank and Sherry A. Glied

Mental illnesses can affect people's moods, cognition, and behavior. As with any illness, costs result from the impairments and functional losses that sickness generates. For mental illness in particular, these costs frequently are also external to the individual, affecting family members and society at large. People with these often-persistent mental illnesses seek recovery—a chance to live full, socially engaged lives—even if a complete cure is not possible. Medical care can help them do this by reducing the magnitude of losses in functioning. Social supports (beyond health insurance) can also defray a portion of the consequences of any remaining impairment.

The evolution of care for mental illnesses has been quite distinct from that of other medical conditions. For other medical conditions, improvements in care have generated large reductions in the functional losses associated with disease, through reduced morbidity and mortality, and have been accompanied by large increases in spending on medical care services. By contrast, treatment advances for mental illnesses have resulted in only modest reductions in functional losses, and costs of treatment for these conditions have only increased at about the rate of GDP growth. Meanwhile, the costs of the remaining impairments associated with mental illnesses have grown rapidly. Despite growth in spending on social supports for this population, people with mental illnesses continue to experience downward economic mobility, and their conditions impose externalities on others.

■ *Richard G. Frank is Leonard D. Schaeffer Chair in Economic Studies at the Brookings Institution, Washington DC. Sherry A. Glied is Dean and Professor of Public Service, Robert F. Wagner Graduate School of Public Service, New York University, New York City, New York. Their email addresses are rfrank@brookings.edu and sherry.glied@nyu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.153.

We begin this essay with background information about the prevalence of mental illness in the United States, along with information on treatments, expenditures, and the social institutions that become involved with mental health issues. We then emphasize that mental illnesses are very heterogeneous, encompassing conditions that have vastly different effects on functioning. For many people with mild to moderate illnesses, recent advances in the quality of mental health treatment could, evidence indicates, alleviate many of the functional consequences of those disorders. Achieving this requires improving the match between clinical problems and treatments (Horvitz-Lennon 2020). For people with persistent relapsing and more serious disorders, treatment can be more difficult, and improving their wellbeing is further complicated by a myriad of socioeconomic factors, including rising housing costs, increased criminalization of disturbed and disturbing behavior, and growing returns to cognitive and interpersonal skills in the labor market. These changes have exacerbated the negative impacts of functional impairments. Improving the well-being of this group will require greater access to more comprehensive social supports. Finally, we offer some thoughts about the implications of these insights for how the United States should be addressing the challenges created by mental illnesses, both through treatment and social supports.

Background

The Prevalence of Mental Illness

Despite decades of research seeking biological markers for mental illness, no definitive laboratory or radiological test yet exists for any mental illness. Thus, estimates of the number of people with mental illness are derived in three main ways: expert or survey-based assessments of symptoms and a determination of whether individuals meet diagnostic criteria for mental illness, survey data in which individuals report whether they have experienced mental health–related functional impairment, and counts of the number of people who have utilized mental health services. The three approaches yield somewhat different estimates in part because they identify groups that overlap, but are not the same.

The first approach focuses on whether individuals meet certain diagnostic criteria. Mental illnesses are diagnosed based on symptoms reported to clinicians or survey interviewers by patients, relatives, or third parties (Kessler 1994). Epidemiologic surveys using structured diagnostic estimates across randomly selected populations find that about 18 percent to 21 percent of the adult population report symptoms that meet the diagnostic criteria for a mental illness each year (Bagalman and Cornell 2018; SAMHSA 2021). Some 3 percent to 5 percent of the adult population meet criteria for what are commonly termed serious mental illnesses, such as schizophrenia, bipolar disorder, and chronic depression. These serious mental illnesses are those most likely to have substantial external effects on family members and the public.

Second, estimates of the prevalence of mental illness can be based on survey questions that ask individuals about mental health–related impairments, such as days spent in poor mental health. About 17.5 percent of the adult population reported having a functional impairment associated with a mental health condition, measured as spending one or more days in poor mental health in the prior month in the National Household Survey of Drug Use and Health. In the Behavioral Risk Factor Surveillance System (BRFSS) survey, among those who reported spending eight or more days in the prior month in poor mental health, 40 percent—twice the share among those with zero days in poor mental health—were socially disengaged; and were not working, self-employed, students, homemakers, or retired (authors' analysis of the BRFSS 2015).

However, an obvious challenge is that some people may report spending days in poor mental health and experiencing functional impairment, but may not meet the full set of diagnostic criteria for an illness. Among the 17 percent who reported spending one or more days in the prior month in poor mental health, 74 percent also responded to questions consistent with a diagnosis of a mental health condition, but the rest did not. Others may experience poor mental health, at least occasionally, but not report it on the survey.

A third approach to looking at the prevalence of mental illness focuses on those who have received treatment for mental illness. In 2019, an estimated 19.2 percent of the total adult population received treatment for a mental health problem (SAMHSA 2021). Nearly half of those 18 and older with a diagnosable mental illness received treatment (including telehealth services) in the pre-COVID quarter of 2020, an increase of about 10 percentage points over the 2008 rate (40.9 percent), which was just slightly higher than the 38.6 percent rate reported in 1985 (Regier et al. 1993). Among those with serious mental illnesses, 65.0 percent received treatment, slightly lower than the rate of 65.7 percent in 2008 (SAMHSA 2021).

Of course, the statement that half of adults with a diagnosable mental illness received treatment also implies that half did not receive treatment. In addition, a substantial share of people using mental health services report relatively low levels of symptoms and distress, levels that do not meet criteria for a diagnosable condition (Germack et al. 2020). These patterns suggest the possibility of both undertreatment of people with mental illness diagnoses and overtreatment of people without diagnoses, but this interpretation may overstate the extent of mismatch. On the undertreatment side, some people may report symptoms of illness over a duration that meets diagnostic criteria, but these symptoms may be insufficiently disruptive to warrant seeking help. In other cases, impairments from mental conditions are present, but demand for appropriate services may be dampened by costs, lack of availability of care, or stigma (Frank and Glied 2006). On the overtreatment side, symptoms that do not meet the specific constellation of criteria required for diagnosis may still cause significant distress. Treatment may be sought to gain assistance in addressing challenges such a navigating the complexities of work, marriage, or family life, rather than because of a need to achieve relief from an illness.

Treatment

There are two main forms of mental health treatment: (1) psychosocial treatments, like psychotherapy and counseling of individuals or groups; and (2) pharmacotherapies, like antidepressants, anxiolytics, antipsychotics, and mood stabilizers, among others. They are delivered in a variety of treatment settings including inpatient, outpatient, residential, and so-called partial hospital programs. Psychosocial treatments and pharmacotherapies are used alone and in combination. In addition, complementary support and counseling for people with serious mental illnesses are sometimes provided by peer counselors, who are people with lived experience of a serious mental illness. Less commonly used are a set of somatic (that is, body related) treatments, particularly electro-convulsive therapy and transcranial magnetic simulation, to treat illnesses such as refractory major depression.

There is strong evidence showing that several of these treatment modalities are effective in reducing symptoms of mental illness, at least in the short-term, for most conditions and many people (Insel 2022). For a review of recent issues in treatment of mental illnesses, see Goldman, Frank, and Morrissey (2020, ch. 5, 11, and 21).

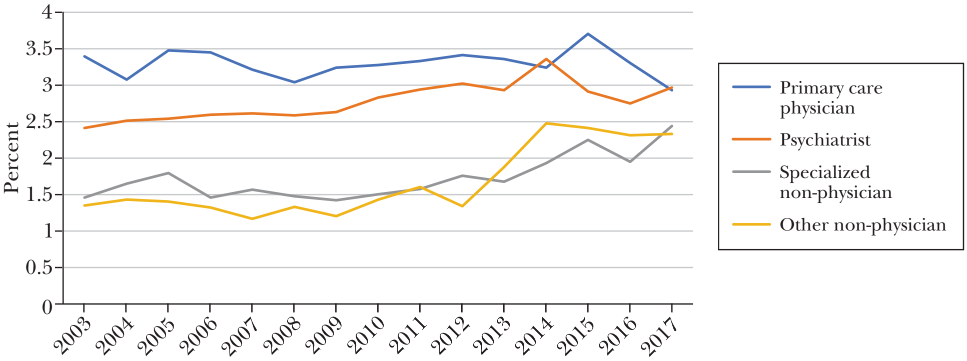
The share of the adult population using mental health treatment has increased over time. The National Health Interview Survey indicates that 19.2 percent of American adults reported receiving some form of mental health treatment in 2019, 15.8 percent used prescription medication for mental illness, and 9.5 percent received psychotherapy from a mental health professional such as psychiatrist, psychologist, psychiatric nurse, or clinical social worker (Terlizzi and Zablotsky 2020).[1]

Figure 1 describes the trends in office-based and outpatient visits for the treatment of mental health conditions by selected provider types, based on data from the Medical Expenditure Panel Survey. In 2017, about 2.9 percent of all adults (36 percent of those with such visits) saw primary care physicians for their mental health care; about one-quarter of those with a visit saw only a primary care physician (not shown on the figure). Some 2.3 percent of all adults (a growing share of those with visits) received at least some care from a nurse practitioner, physician assistant, or other nonspecialized nonphysician providers. Another 3 percent (36 percent of all those with visits) received care from a psychiatrist.

The proportion of adults receiving care from a specialized mental health professional who is not a physician, such as a psychologist or social worker, has also increased substantially over time and now accounts for about 30 percent of all those with visits. Specialized nonphysician providers provide the bulk of psychotherapy services, which typically require multiple visits. In 2017, physicians, including psychiatrists and primary care doctors, provided 43 million mental health–related

[1] Data on expenditures among those with mental health conditions show similar patterns. An estimated 16 percent of American adults had such expenditures; about 14 percent of adults had expenditures for mental health related prescription medications and about 8 percent had either an office-based or outpatient mental health-related visits in 2017 (based on authors' analysis of data from the Medical Expenditure Panel Survey, available at https://www.meps.ahrq.gov/).

Figure 1
Trends in Adult Mental–Health Related Office and Outpatient Visits by Provider Type



Source: Medical Expenditure Panel Survey (2003–2017).
Note: Adults aged 18 and older with an office-based or outpatient visit that includes a diagnosis classified as "Mental, Behavioral and Neurodevelopmental disorders" (ICD9: 290-319 or ICD10: F01-F99).

psychotherapy or counseling visits, while nonphysicians, including psychologists, social workers, and mental health counselors, provided 99 million such visits. There are persistent differences in the use of psychotherapy by race, ethnicity, income, and educational status. For example, rates of use of psychotherapy for non-Hispanic whites are 35 percent higher than that of Blacks and 65 percent higher than the rate for Hispanics (Terlizzi and Zablotsky 2020).

Most people with mental illnesses receive all their treatment in outpatient settings, but a small minority are hospitalized for short or long periods of time, either by choice or through involuntary commitment. Roughly 0.7 percent of the population has a hospital admission with a primary mental illness diagnosis and 2.6 percent has either a primary or secondary diagnosis of a mental illness (Owens et al. 2019).

Expenditures on Treatment

Total spending on treatment services for mental illnesses was estimated at $180 billion in 2016 (Dieleman et al. 2020) and projected to be $238 billion for 2020, or 1.1 percent of GDP (SAMHSA 2014).[2] In 1975, the mental health services share of GDP was similar, at about 1 percent. In this way, mental health spending growth differs strikingly from overall health care spending growth,

[2] Note that the government's National Mental Health Expenditure Accounts have not been produced since 2014. At that time projections were made onto 2020. Dieleman et al. (2020) produced expenditure estimates for 2016 based on spending and service utilization data reported for that year.

which increased from 6.5 percent of GDP in 1975 to 19.7 percent in 2020. This difference is also notable because, as Figure 1 shows, the share of the population receiving mental health treatment has increased over this period. There are four main reasons for the difference in spending growth between mental health care and general medical care.

First, the main driver of cost growth in the general health care sector has been technological change, particularly through the introduction of capital-intensive devices and procedures (Chernew and Newhouse 2011). In contrast, the technology of treatment in mental health continues to rely on labor and prescription drugs. Newer treatments for mental health conditions have typically offered few gains in efficacy, although they have generated improvements in treatment adherence and outcomes by reducing side effects and increasing the tolerability of treatments (Insel 2022). While psychopharmacology experienced considerable innovation prior to 2000, relatively few new classes of drugs for treating mental illnesses have been introduced since then. The sluggishness in innovation in psychopharmacology stems in part from the complexity of the characteristics of mental illnesses and the ways that psychotropic agents work in the body (Drukarch, Jacobs, and Wilhelmus 2020). A variety of proposals have been made to address how to increase the pace on innovation (Insel 2022). Some suggest encouraging research in directions that recognize the complexity of but run counter to the prevailing molecular approaches to drug development (Drukarch, Jacobs, and Wilhelmus 2020). Others focus on the characterization of illness, and still others take aim at the regulation of the research enterprise (Tyrer 2009). In addition to the slowdown in the introduction of new drugs, as existing major psychotropic medications lost market exclusivity and faced competition from generic equivalents, prices for these medications have declined. Along with changes in the psychopharmacology market, there have also been evidence-confirmed improvements in psychosocial treatment, such as the refinement of cognitive behavioral psychotherapy. These have affected the content of talk therapy, but have typically reduced (or not increased) the time required and the cost of labor needed to deliver the service. Although the evidence base for psychosocial treatment has expanded considerably, few practitioners today consistently use evidence-based psychosocial treatments (Areán and Ratzliff 2020).

Second, over the past 50 years there has been dramatic, cost-reducing substitution for the human and institutional inputs that were previously used to provide mental health care. In 1975, 63 percent of mental health care spending was for institutional care in hospitals and nursing homes; today, 31 percent of expenditures occur in these costly settings (SAMHSA 2014; 2016). Treatment with prescription drugs has taken a central position in treatment of mental illnesses, often substituting for costlier psychotherapy for the most prevalent mental health conditions, depression and anxiety. In 1971, spending on psychotropic drugs amounted to 4 percent of spending on treatment of mental illness, but that share doubled by 1986 and reached 26 percent in 2020 (SAMHSA 2016; Levine and Levine 1975).

The cost of psychotherapy itself has also dropped sharply because the mental health sector has been far more accommodating of diverse types of health care providers than has general health care. Psychotherapy provision has shifted from treatment by psychiatrists and PhD-level psychologists to treatment by social workers, counselors, and MA-level psychologists. In 2021, psychiatrists earned a median salary of $327,000, PhD-level psychologists earned a median of $179,000, MA-level psychologists earned $81,000, social workers earned $57,800, and counselors earned $48,500 (Bureau of Labor Statistics 2022). Today over 90 percent of psychotherapists are trained below the doctoral level, a far higher share than in the 1970s and 1980s. The shift towards lower cost professionals with less extensive training has driven the costs of psychotherapy down, without any documented evidence of a reduction in quality—although no recent studies have directly compared the quality of services delivered by those with varied professional training. An older literature found similar outcomes for psychotherapy across disciplines (Office of Technology Assessment 1980). The widespread dissemination of mental-health services delivered via web-based or smart-phone apps may reduce these costs even further.

Third, a much larger share of mental health care (just under two-thirds) is paid for by public funds (about one-third is paid by Medicaid) than is the case for general health care, and a much larger share—20 percent—is paid for by programs under fixed budgets. Public programs generally pay lower prices. For example, nominal spending per service user for state-funded mental health care increased about 12 percent from 2008 to 2019, slower than economy-wide inflation (based on our own calculations from funding reports by the National Association of State Mental Health Program Directors). Thus, the difference in payment sources and rationing arrangements that govern mental health services tended to reduce growth rates in mental health spending.

Finally, mental health spending appears to be growing much more slowly than general health spending, in part because of a change in classification. In the 1970s and 1980s, when institutional treatment of those with serious mental illness accounted for a much larger share of mental health spending than it does today, all the expenses of institutional treatment—including the costs of whatever limited clinical treatment was provided as well as the costs of institutional room and board, often of poor quality—were counted as part of mental health spending. Today, the costs of housing and food for people with serious mental illness, who are not typically institutionalized, are no longer counted as part of mental health treatment spending.

This classification change hints at a deeper problem in accounting for the costs of mental illness. Mental illnesses can be functionally disabling and can lead people to behave in ways that generate negative externalities. These features of mental illness explain why there was substantial public involvement in the support and control of people with these conditions for centuries before any effective forms of treatment existed. The primary functions of the institutions that accounted for so much of mental health spending before 1970 were to ensure that people with the most serious illnesses received food and shelter—and that they did so out of sight of the rest of society (Grob 1994). Today, a wide variety of public services and institutions play some of the same roles, including the criminal justice system, housing and homeless services, disability and other income support programs, and specialized educational and employment services.

Social Programs, Institutions, and Support of People with Mental Illnesses

Societies have always struggled with how to address the impairments and externalities associated with serious mental illnesses, which can often result in various forms of self-harm, including self-neglect, risky behaviors, and even suicidality (Mechanic, McAlpine, and Rochefort 2014). These illnesses are also associated with a range of external effects: family dissolution, homelessness (Leopold 2020), crime and crime victimization (Osher and Thompson 2020; Swanson, Barry, and Swartz 2020), and behaviors that are both frightening and stigmatized. Until about 50 years ago, societies mainly addressed these problems by involuntarily hospitalizing people with serious mental illnesses, thus removing them from their families and communities and providing them with custodial care (though even at the height of institutionalization, many people with serious mental illness were not institutionalized and received no services at all). In the early 1970s, 57 percent of total mental health expenditures were paid by state and federal government (Levine and Levine 1975). State-funded care consisted largely of direct funding of custodial psychiatric institutions that served the sickest patients and those from low-income families. Since then, therapeutic, legal, and financial changes have all contributed to a substantial movement of this population out of institutional care (Grob 1994; Mechanic, McAlpine, and Rochefort 2014). The current pattern of financing mental health care, through private insurance, Medicare, and Medicaid, more closely resembles that of general health care. Currently, only 20 percent of mental health spending comes in the form of direct government funding of providers.

The first antipsychotic drugs were introduced in the late 1950s (Ban 2007), offering the possibility that some of the symptoms of serious mental illness could be at least partially controlled outside the hospital. Beginning about a decade later, a series of court cases codified the conditions under which someone could be involuntarily committed to psychiatric care and established affirmative rights to treatment, the right to refuse treatment, and the right to be treated in the least restrictive setting (Grob 1994; Mechanic, McAlpine, and Rochefort 2014). In addition, the introduction of the Medicaid program in 1965 offered states the opportunity to shift part of the cost of care for people with mental illnesses to the federal government—but under program rules, only if that care was not provided in state psychiatric hospitals. The confluence of these forces led to rapid deinstitutionalization of mental health treatment through the 1970s and 1980s.

However, even successful treatments for serious mental illness like antipsychotic drugs were not effective enough to allow people with these illnesses to live outside institutions without any further support. Thus, the Supplemental Security Income program in 1974 and expansions of eligibility for the Social Security Disability Income program created new opportunities for people with disabling mental illnesses to be financially sustained outside of total institutions. Over time, additional specialized employment and housing support programs (like "Section 8" vouchers), as well as expansion of mainstream programs such as food stamps (now officially the Supplemental Nutrition Assistance Program) have supplemented these income support programs.

Table 1 summarizes the size of the population with serious mental illnesses that use each type of program and the estimated spending associated with their support. A detailed description of how the estimates were constructed is available in the online Appendix.

The disabling nature of mental illnesses often means that simply making treatment and services available is not enough to ensure that people in need are served. Effective treatment responses for serious mental illnesses today require combining clinical treatment elements with high levels of contact and patient engagement and include help with navigating their community environments like housing, income support enrollment, and adherence to medication regimens. Ideally, these combined functions provide key supports that were previously supplied by psychiatric hospitals, but with improved treatment and much greater deference to patient autonomy. In practice, some people with serious mental illness do not receive or benefit from the full complement of services and remain indigent, disabled, and socially excluded. This group constitutes a disproportionate share of those in homeless shelters and the criminal justice system.

Combining estimates for medical spending on the treatment of mental illnesses and these estimates of the cost of social supports suggests that the cost of these components together amount to about $320 billion, with about two-thirds of this cost consisting of medical care costs and the remaining one-third consisting of direct social services and income supports. Recent estimates of similar costs from Australia (using methods like ours above) and Denmark (using a registry-based approach) both find about one-third of the total cost of treatment and social support for people with mental illness is for treatment and two-thirds is for social support, reversing the US proportions (Productivity Commission of the Australian Government 2020; Christensen et al. 2022),[3] with 30 to 40 percent lower total per capita spending rates overall. Part of this differential is due to the much higher costs of healthcare in the United States than in either Australia or Denmark; US per capita healthcare expenditures average about double those of these other countries. Justice-related costs are also much higher in the United States than in these other countries, reflecting much higher overall rates of incarceration in the United States (160 and 68 per 100,000 in Australia and Denmark versus 639 per 100,000 in the United States [https://www.prisonstudies.org/about-us]). By contrast, expenditures

[3] The Australian study finds spending of $11.3 billion (in Australian dollars) for medical care services, $4.1 billion for direct services, and $10.9 billion for income support, reflecting a system with a much greater focus on nonmedical services. Danish estimates from a registry-based system find €1.2 billion for medical care services and public transfers of about €2.2 billion.

Table 1

Spending for Non–health Care Support and Services for Mentally Ill People in 2019

	With Indicators of Mental Illness	
Program type	*Number of people*	*Spending*
In the general population . . .		
Housing and Urban Development (HUD) housing assistance	963,000 households	$8.84 billion
Supplemental Nutritional Assistance Program (SNAP)	3.74 million households monthly	$11.2 billion
Temporary Assistance for Needy Families (TANF)	208,000 families monthly	$1.11 billion
Supplemental Security Income (SSI)	1.61 million	$11.7 billion
Social Security Disability Insurance (SSDI)	2.39 million	$31.2 billion
Total		$64.1 billion
People experiencing homelessness . . .		
Grants for services	insufficient data	$498 million
People involved with criminal justice system . . .		
Local jails	196,000 daily	$6.89 billion
State prisons	178,000	$6.16 billion
Federal prisons	13,800	$489 million
Total		$13.5 billion
Police responses	insufficient data	$12.3 billion
Total from all programs		$90.5 billion

Sources: Percent of people receiving HUD housing assistance, SNAP, and TANF who had indicators of mental illness determined from 2019 National Health Interview Survey data using the GAD-7 and PHQ-8 scales (authors' analysis of NHIS data). Number of households in subsidized housing and average expenditure per household from 2019 HUD Picture of Subsidized Housing data; number of households receiving SNAP monthly and average expenditure per household from Supplemental Nutrition Assistance Program 2019 National Level Monthly Summary; number of families receiving TANF monthly and average expenditure per family from the 2019 Characteristics and Financial Circumstances of TANF Recipients (HUD 2020; USDA 2020; HHS 2020).

People with indicators of mental illness receiving SSI and SSDI based on recipients who had severe persistent mental illness, listed in one of the following diagnostic categories: Schizophrenic and other psychotic disorders; Mood disorders; Organic mental disorders; Other mental disorders. Number of people with serious and persistent mental illness (SPMI) and average annual spending per person from 2019 SSI Annual Statistical Report and the Annual Statistical Report on the Social Security Disability Insurance Program (SSA 2020a, 2020b).

A survey of the homeless is the 2019 Point-in-Time count, completed by Continuums of Care (CoC) (HUD 2019), but this survey has no standardized diagnostic interview or screening instrument for mental illness, and so no estimate is given for number of people in this category. For the estimate of annual spending for the homeless mentally ill, most federal funding for homelessness is distributed through CoC grants and Emergency Solutions Grants (Lucas 2017). The spending number in the table was calculated from the sum of these grants in 2019 times the fraction of people experiencing homelessness classified as severely mentally ill from the PIT counts (HUD 2018a, 2018b).

For calculations relevant to incarcerated people, prevalence of mental illness among people in jail from the Bureau of Justice Statistics's Indicators of Mental Health Problems Reported by Prisoners and Jail Inmates, based on the National Inmate Survey (NIS-3) from 2011-2012 (Bronson and Berzofsky 2017). Prevalence of mental illness among people in prisons from the BJS's "Indicators of Mental Health Problems Reported by Prisoners," based on the 2016 Survey on Prison Inmates (Maruschak and Bronson 2021). The number of incarcerated people in jail is the average daily population of people in jail from the BJS's "Jail Inmates in 2019" report (Zeng and Minton 2021). The number of incarcerated people in prison is taken from the number of people in prison at year end in 2019 from the Prisoners in 2020 Statistical Tables (Carson 2021). Spending per inmate in jail is the annual cost per inmate in 2017 inflated to 2019 dollars (Horowitz 2021). Spending per inmate in state prisons is the annual average per inmate from the National Institute of Corrections' data on 2019 National Averages (NIC 2019). Spending per inmate in federal prison on average in 2019 is taken from the Federal Register's Annual Determination of Average Cost of Incarceration Fee (BOP 2021).

Numbers on mental illness among people involved in police responses are inconsistent and limited by lack of data collection. Police spending on people with mental illness based on research from the Treatment Advocacy Center (2019), which found that 10 percent of law enforcement budgets were spent on calls involving people with mental illness, and research from the Urban Institute (2020), which found that state and local government spent $123 billion on police in 2019, using data from the US Census Bureau Annual Survey of State and Local Government Finances.

Notes: Figures are annual estimates for the year of 2019 unless otherwise noted. For details on calculations, see online Appendix.

on housing and homelessness services, employment support, and income support for those with mental illness are much higher in these comparison countries.

These differences raise the possibility that accounting for savings "across buckets" might generate a stronger fiscal case for mental health interventions—that is, improvements in the quality of social supports provided to people with mental illness might reduce the clinical costs associated with treatment or that better treatment could reduce the costs of social support (Newman et al. 1994). To date, however, there is limited empirical evidence that expansions of either treatment or social support spending can substantially reduce public spending in other buckets. A small number of studies have estimated costs across buckets for a specific population, which requires tracking a population across disparate administrative systems. A still smaller number have done this in the context of an intervention. These few studies (of mental health courts and housing first interventions) have not found reductions in system-wide costs associated with these interventions (Steadman et al. 2014; Ly and Latimer 2015). The existing evidence suggests that these programs may be complements, in the sense that mental health treatment offered to people with adequate social supports works more effectively than treatment offered to those with inadequate support.

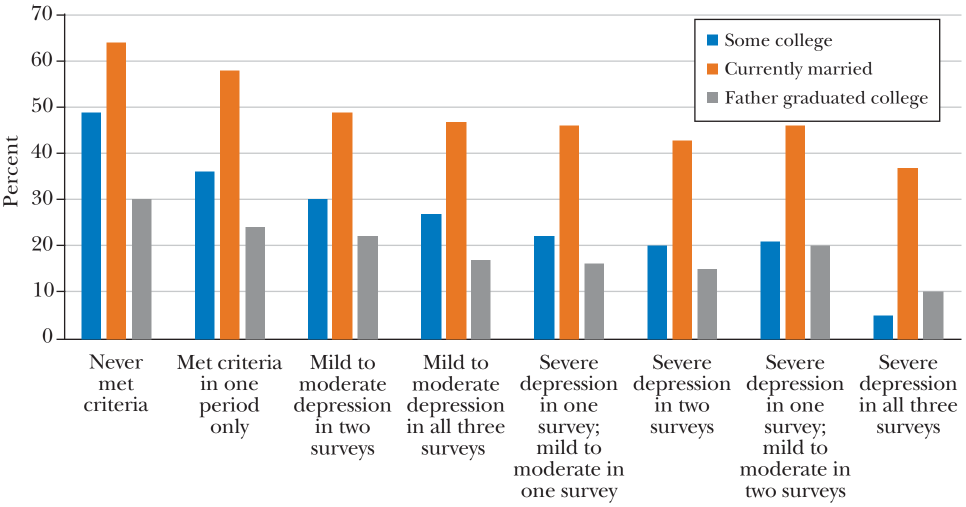
Heterogeneity in Mental Illness and in Its Economic Effects

All mental illnesses cause a loss in well-being, by definition, and most have at least transitory economic effects. Because mental illnesses are disproportionately likely to manifest first in adolescence or early adulthood, they can also have consequences on the accumulation of human capital. The nature and magnitude of the consequences of mental illness, however, vary tremendously across the broad range of conditions encompassed within this category.

Longitudinal information on the well-being of the overall population with mental illness is quite limited. In Figure 2, we use the National Longitudinal Survey of Youth 79 (NLSY79) to provide baseline demographic information on the population classified into eight categories describing the severity and persistence of the most prevalent of the major mental illnesses, depression. We construct the groupings according to scores on three depression screening surveys (conducted in 1992, when the participants in the sample were at age 27–34, at age 40, and at age 50). Thirty-four percent of the NLSY79 sample met diagnostic criteria for depression in at least one of the three surveys conducted over this 16–23 year period. Among those who met criteria at least once, one-third met criteria on more than one survey.

The striking pattern in this figure is that even those who met criteria for depression in only one of the three surveys are disadvantaged relative to those who never met criteria for depression. While the sample sizes for those with recurrent and severe illness are small, those with persistent and severe depression are much less likely to attend college (48 percent among those with no subsequent depression

Figure 2
Characteristics of People with Different Life Course of Depression

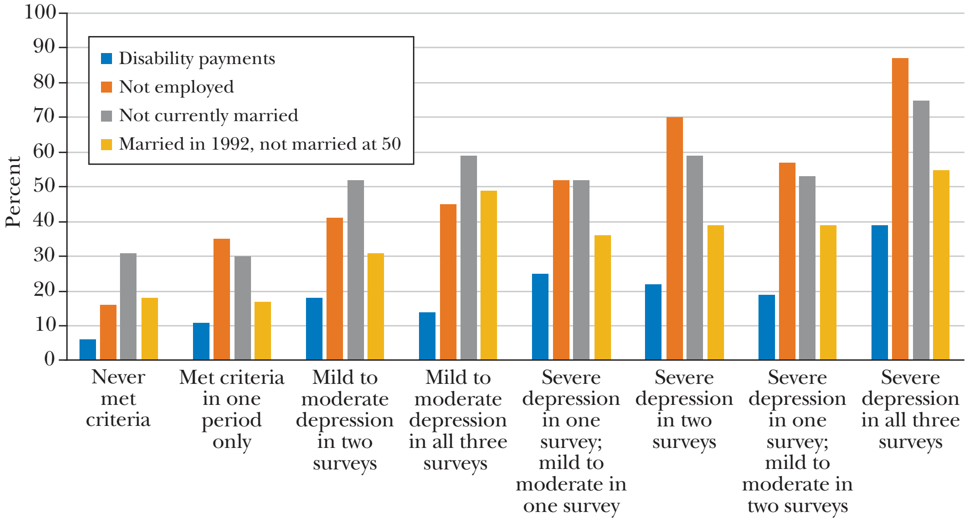


Source: National Longitudinal Survey of Youth 79 (NLSY79) 1992–2016.
Note: Authors' analysis of NLSY data. Depression measured in 1992, at respondent age 40, and at respondent age 50. Depression characterized based on CES-D 7-item cutoffs at 0–7 (none), 8–13 (mild to moderate), and 14–21 (severe). See online Appendix for details.

versus 5 percent among those who had severe depression in three surveys), to be married when they are aged 27–34 (64 percent versus 37 percent), or to have grown up in a family where the father graduated college (30 percent versus 10 percent) than their counterparts who do not have subsequent depression episodes.

Figure 3 describes selected dimensions of the circumstances at age 50 for the same groups shown in Figure 2. Again, substantial differences are displayed between those who never met criteria and those who met criteria even in one period only. Increasing severity of illness is correlated with much higher rates of disability payment receipt (6 percent among those with no depression episodes versus 39 percent with three serious episodes), much higher rates of nonemployment (16 percent versus 87 percent), higher rates of not being married (32 percent versus 83 percent), and much higher rates of marital instability (20 percent versus 55 percent) (consistent with earlier studies, including Anesetti-Rothermel and Sambamoorthi 2011; Danziger, Frank, and Meara 2009; Kessler, Walters, and Forthofer 1998). As we discuss below, a two-way relationship seems plausible here; that is, mental illness causes poorer life outcomes and poor life circumstances increase the risk of subsequent mental illness. However, the evidence presented here supports our arguments that there is a correlation between even transient episodes of mental illness and decrements in well-being. Moreover, varying severities of mental illness appear to be associated with very different lifetime outcomes.

Figure 3
Characteristics at Age 50 by Lifetime Depression Trajectory



Source: National Longitudinal Survey of Youth 79 (NLSY79) 1992–2016.
Note: Authors' analysis of NLSY data. Depression measured in 1992, at respondent age 40, and at respondent age 50. Depression characterized based on CES-D 7-item cutoffs at 0–7 (none), 8–13 (mild to moderate), and 14–21 (severe). See online Appendix for details.

Serious and Persistent Mental Illness

Having a serious and persistent mental illness is more common with some diagnoses than with others—schizophrenia and bipolar disorder, for example—but there is considerable heterogeneity in the level of functioning within these diagnostic groupings. People with serious and persistent mental illnesses have very low rates of educational attainment, labor force participation, and family formation, and are more likely to subsist on disability income payments through Social Security Disability Insurance and to receive clinical services paid for by Medicaid. Efforts to increase labor force participation among people who have already met criteria for disability payments, through supported employment programs, have been only very modestly successful and do not show promise in reducing reliance on disability payments (Glied, Frank, and Wexler 2020; SSA 2022; Bond et al. 2015).

Serious and persistent mental illnesses may affect people's ability to care for themselves and may lead them to behave in ways that are disturbing to others. The infrastructure of care and treatment for those with such conditions has evolved to be much more proactive—or intrusive—than conventional treatments and social supports. We have previously estimated that about 29 percent of those in this group were institutionalized in psychiatric hospitals at the height of institutionalization.

That figure has fallen to below 5 percent.[4] Nonetheless, all states operate systems of involuntary commitment that allow them to hold in custody people who pose a danger to themselves or others. The US Supreme Court decision in *Olmstead v. L.C.* (527 US 581 [1999]) requires that such persons be treated in the least restrictive environment possible for their conditions. The gold standard treatment of people with such conditions usually involves the use of teams of professionals, such as Assertive Community Treatment teams. These are multidisciplinary teams who conduct outreach, provide treatment, and offer community supports for the comprehensive set of needs of a small group of severely-ill individuals living in the community. Evidence from multiple randomized controlled trials shows that the use of such teams can reduce inpatient hospitalization rates (Scott and Dixon 1995). Providing people with serious and persistent mental illness with supported housing—a unit combined with caseworker monitoring and support—can improve housing stability and disability. A growing body of evidence also suggests that connecting people to intensive services in the early stages of serious illnesses such as schizophrenia or bipolar disorder can moderate the trajectory of these conditions (Kane et al. 2016). The highly specialized and intrusive nature of these services suggests that there is relatively little moral hazard in making access to them readily available to those in need, as those likely to achieve little benefit from these services would be unlikely to seek them even at a very low price.

People with serious and persistent illnesses can live better lives when they have access to the full range of evidence-based services and disability supports, though in most cases, these treatments and services do not "cure" the underlying disease. Unfortunately, access to these services is highly incomplete due in part to supply limitations, policy designs, and features of the illnesses themselves. For example, receipt of benefits through Social Security Disability Insurance requires having an established work history and navigating a substantial and multistep application process. This process is intended to reduce moral hazard in receipt of these services, but it is quite imperfect. For example, Danziger, Frank, and Meara (2009) found that among low-income women whose applications for Supplemental Security Income benefits were rejected, over one-third met diagnostic screening criteria for serious mental illness, and rates of housing instability in this rejected group were nearly 50 percent higher than among those who did receive benefits.[5] Even when people, or their proxies, are able to navigate these eligibility processes, there are not enough supported housing units, Assertive Community Treatment program slots, or other evidence-based services available to care for all those with these serious mental illnesses.

[4] Authors' calculation based on utilization rates for people with SMI from NHSDUH applied to National Mental Health Services Survey (N-MHSS): 2018 Data on Mental Health Treatment Facilities

[5] Supplemental Security Income (SSI) and Social Security Disability Insurance (SSDI) both require impairments that prevent engagement in Substantial Gainful Activity. SSDI requires a work history of 40 quarters to qualify for benefits. SSI has the same functional impairment standard but does not require a work history to receive benefits. SSDI payments are higher than those from SSI.

A variety of changes in the broader environment have also contributed to declines in the well-being of this most vulnerable group. People with serious and persistent mental illness were disproportionately swept up in the rise in jail and prison incarceration beginning in the 1980s. Some 26 percent of people in jails and 14 percent in prisons have a serious mental illness (Bronson and Berzofsky 2017). While there is no evidence of a secular increase in the fraction of the incarcerated population that has a mental illness, and estimates suggest that only a very small fraction of the rise in incarceration reflects deinstitutionalization, the overall rise in incarceration over time meant that more people with severe mental illnesses were incarcerated (Raphael and Stoll 2013; Frank and Glied 2006). People with severe mental illness are also disproportionately affected when police are assigned the lead role in controlling disturbed and disturbing behavior. Analysis of the *Washington Post* database on police shootings shows that 23 percent of the 6,800 police shootings since 2015 have involved people with mental illness.[6] Increases in city rents, as well as changes in housing codes that have significantly reduced the availability of very cheap housing options, have also contributed to increases in the share of people with mental illness who are homeless (O'Flaherty 1998; Frank and Glied 2006). This problem is exacerbated by the fact that monthly Supplemental Security Income payments, which are a principal source of income from which people with disabling mental illnesses pay rent, are indexed only to the general rate of inflation—and thus have not kept up with rising rental costs.

Moderate Recurrent Mental Illness

Many moderate mental illnesses are chronic, with early onset and episodic relapses. For example, the average age of onset of depression is 24, and the recurrence rate of this condition is about 80 percent (Sim et al. 2015). Moderate mental illness often affects workplace-relevant skills, including increased absenteeism, along with difficulties in interacting with customers and coworkers and concentrating on the job (Millan et al. 2012; Rock et al. 2013; Adler et al. 2006).

Early onset illnesses are also likely to affect human capital accumulation. People with early onset depression (particularly women) are less likely to obtain college degrees and less likely to pursue postgraduate degrees (Berndt et al. 2000). They are more likely to enter occupations that require fewer cognitive and interpersonal skills, and that have lower subsequent wage growth (Wang, Frank, and Glied 2022). In a longitudinal study using Finnish registry data, Hakulinen et al. (2019) find that people aged 15–25 with depressive disorders are three times more likely to become unemployed and have much lower educational attainment and wages at age 50 than their counterparts who did not have depression. Fletcher (2013), using data from the National Longitudinal Study of Adolescent to Adult Health (often called Add Health), finds that adolescent depression decreases the probability of employment in the late-20s by 5 percentage points and reduces income

[6] Authors' calculations based on a *Washington Post* data base that can be queried and is available at: https://www.washingtonpost.com/graphics/investigations/police-shootings-database/

at that age by 15 percent. People with major depression at baseline in the Health and Work Study sample, a group recruited from physicians' offices in Massachusetts, are four times more likely to change occupations and six times more likely to become unemployed than those without this condition at six-month follow up (Lerner et al. 2004). Recurrent illnesses also affect human capital accumulation through their effects on employment, hours, and absenteeism. In the National Longitudinal Survey of Youth, by age 50, people with recurrent depression had accumulated 418 fewer weeks of work experience than their counterparts without depression (authors' calculations from the NLSY79).

Many evidence-based mental health treatments, such as work-focused cognitive behavioral therapy, have been shown effective in reducing symptoms and improving health for workers with recurrent depression and other moderately severe mental illnesses at a point in time (Lerner et al. 2012; Lerner et al. 2015; Cullen et al. 2017). A few studies have also examined the impact of these treatment on employment and workplace outcomes. Those studies find that treatment-induced reductions in depressive symptoms are associated with reductions in work impairments and improvements in work outcomes, though the clinical changes are larger than the labor supply responses (Lagerveld et al. 2012; Kröger et al. 2015).

Improvements in functioning at a point in time are important, but reducing the well-being losses associated with recurrent illness also requires continuation of treatment and monitoring. Even those who receive evidence-based treatment remain at risk for future work disruption, labor force exit, and unemployment—and potentially, the loss of health insurance coverage (Lerner et al. 2012). They may also face barriers to reestablishing relationships and good work habits and may face workplace stigma after an episode of illness. Additional interventions within the workplace may be needed to maintain productivity and retain workers.

One potential policy lever for promoting greater use of workplace interventions is the Americans with Disabilities Act of 1990, which requires employers to offer reasonable workplace accommodations—including for disabilities related to mental illness. For people with mental illness, job coaching is the most frequently used accommodation, as well as switching from full-time to part-time work hours (Granger, Baron, and Robinson 1997). Workers with mental illnesses have historically been almost 50 percent less likely than other workers seeking accommodation to receive them (Zwerling et al. 2003). In part, the pattern occurs because of the ambiguity surrounding what constitutes a reasonable accommodation and which accommodations will most effectively help employees with mental health impairments do their jobs (Hickox and Hall 2018). It can also be difficult to design accommodations for episodic mental illnesses when the types and level of impairments are not consistent.

Mild or Nonrecurrent Moderate Mental Illness

Most people who ever experience symptoms consistent with a diagnosable mental illness will not experience frequent recurrences or debilitating functional impairments. The very high prevalence of these conditions, however, suggests that they impose significant social costs.

For the population with single episodes of less severe mental illness, a number of existing evidence-based treatments have been shown to be effective, including cognitive behavioral therapy, dialectical behavioral therapy, medication treatment, and other therapies for depression, anxiety, obsessive-compulsive disorder, attention-deficit disorder, and other conditions (for a listing, see https://www.samhsa.gov/resource-search/ebp). Unfortunately, much of the time, people receive inappropriate or inadequate treatment, even for conditions for which the evidence base is strong (for a detailed review, see Horvitz-Lennon 2020). The challenges in diagnosing mental illnesses and in accurately describing the treatment received, especially for nonpharmacological treatment, have slowed progress in improving the quality of treatment. It is difficult and costly to assess new mental health treatment for many reasons: because mental health does not manifest as a readily observable lesion, because it is difficult to ensure that therapists adhere to specific therapies, and because placebo treatment, which provides people with attention and reassurance, often has substantial effects.

Episodic mental illnesses affect functioning and have economic costs, but the risk of such illnesses can also be heightened by economic difficulties and reduced through economic supports (including health insurance). For example, increases in the earned income tax credit (EITC) reduce poor mental health days (Evans and Garthwaite 2014), and increases in the minimum wage reduce depression and suicidality (Reeves et al. 2017; Dow et al. 2020). Unconditional cash transfers in childhood reduce the risk of anxiety and depression symptoms, and this effect persists into adulthood (Costello et al. 2003; Copeland et al. 2022). In Oregon, people who won a lottery to gain Medicaid coverage saw reductions in depression symptoms even before any gains in treatment might have taken effect. Layoffs and plant closings, conversely, raise the risk of mental illness (Brand, Levy, and Gallo 2008). The mental health effects of income gains and losses suggest that the welfare implications of business cycles and of redistributive programs may be greater than standard analyses might otherwise suggest.

Implications for Economic Research and Policy Analysis

Models of clinical treatment and social insurance coverage for mental illness are often structured around single episodes of care, stable health conditions, and participation that is initiated and maintained by the beneficiary of the treatment. Mental illness is, however, often an ongoing condition with episodic fluctuations, and the illness itself can make it difficult for the patient to initiate and continue coverages. The historic structure of US medical care—employment-based coverage of fee-for-service individual visits—fit particularly poorly with common features of mental illness. The effects of mental illness on maintaining employment meant that a disproportionate share of people with mental illness were uninsured prior to expansions of the Patient Protection and Affordable Care Act of 2010 (Frank, Beronio, and Glied 2014). The use of cost-sharing as a method of rationing people's willingness to seek care was a main form of benefit design prior to the passage of legislation mandating parity in insurance coverage for mental health and physical illnesses in the mid-2000s. Likewise, rationing via cost-sharing disadvantaged many lower-income people needing care. Medicaid, which provides coverage to a disproportionate share of people with mental illness who often qualify because they have low incomes, pays low rates to providers, which limits supply and thereby access to services.[7] These disjunctions have implications for appropriate treatment and social policies related to mental health.

The different levels of mental illness suggest different levels of policy response. For serious mental illnesses, the associated internal harms and external costs associated suggest that we should be encouraging people to use supported housing and increasing the supply of case workers and employment services. For example, this could include expansion of supported housing through the addition of housing vouchers for people with disabilities under the Department of Housing and Urban Development's Section 8 voucher program. Improving the availability of and access to such services broadly would likely improve the well-being of individuals, mitigate external effects, and benefit society broadly (Keisler-Starkey and Bunch 2021; Saunders and Rudowitz 2022). Shifting the first response for disturbing behavior away from the criminal justice system and toward the mental health system can plausibly help to reduce individual and social costs (Tsai 2021). Assertive Community Treatment programs that intensively manage the care of people with serious mental illnesses in community settings have been shown in randomized trials to reduce hospitalizations and stably house people, thereby reducing homelessness and victimization (Latimer 1999; Stein and Santos 1998). The evidence on the impact of ACT on the incarceration of people with serious mental illnesses is relatively weak; one randomized trial showed some evidence of reduced rates of re-arrest (Cuddeback, Simpson, and Wu 2020).

Expanding the availability and generosity of social and income supports does raise concerns about moral hazard: that is, the risk of people being diagnosed as mentally ill to qualify for expanded benefits. However, moral hazard is unlikely to be a substantial problem for in-kind benefits tailored to people with severe mental illness—such as supported housing, supported employment, and care management—because the nonpecuniary dimensions of participating in such programs, which involve a high degree of monitoring, are likely to deter those with lower levels of need. Moral hazard is more of a concern for income support programs, but the extant evidence suggests that the current strategies that link mental illness to income support programs are overly stringent, in the sense that they require an extreme level of mental impairment before conferring eligibility for

[7] An estimated 22 percent of people with mental illnesses are covered by Medicaid compared to 17.8 percent of the population overall in 2020 (Saunders and Rudowitz 2022; Keisler-Starkey and Bunch 2021).

such programs. As a result, people with serious mental illnesses who would quality for such support under a more reasonable standard instead end up overrepresented among those who experience extreme poverty.

For people with less severe mental illnesses, it would be useful to avoid financial incentives to ration the use of mental health care; instead, expansions of Medicaid to include people with mental health conditions who might not meet, or seek, disability determinations may enable better access to treatment. Shifting from price-based rationing to more administrative rationing designs such as those known as managed care is one possible approach. This would mean payments would be made under risk-based performance contracts informed by large population-based spending and performance data. Such arrangements may improve the match between need and treatment (though there is limited evidence of this to date). The clinical flexibility afforded under such payment arrangements may encourage the implementation of evidence-based continuation and maintenance care, which can reduce the likelihood of recurrent episodes (Blier et al. 2007).

In addition, for those with less severe levels of mental health care, policymakers need to be mindful of the relationship between work and illness and, in particular, to consider how to extend the scope of workplace accommodations to encompass mental illnesses. As one example, the rise of remote work offers new opportunities to allow people to be productive with lower levels of interpersonal engagement and stress. Similarly, training of supervisors in support of people with mental illnesses and at work coaching has been shown to improve work stability for people with depression and other mental illnesses (Lerner et al. 2012). At the policy level, the two-way relationship between work and mental illness suggests caution in imposing work requirements on receipt of Medicaid benefits or income support. After all, losing employment, income, or health insurance can trigger mental illness, and mental illness, even at levels well below those required for a disability determination, can make it difficult to obtain and maintain employment.

Conclusion

Mental illnesses are costly and often highly stigmatized and generate notable externalities. They are also heterogeneous, often in ways that are hard to observe, both in their direct effects and in their responsiveness to treatment. Thus, it has proven difficult to craft appropriate policy responses for treatment and social supports.

Current policy choices have led to a misallocation of resources in the delivery of clinical services. Too few people with treatable mental health conditions, including those with serious illness, obtain care that could help them. This situation may arise, in part, because the decisions of people suffering from mental illness to seek care may not accurately reflect the likely value of such care to themselves and to others, as well as because of underinvestment in treatment capacity for the most serious conditions. At the same time, moral hazard associated with insurance coverage of mental health services may lead to overuse (or inappropriate use) of some services within this category, either to address problems of living that cause relatively little impairment or because the quality and nature of treatments are so variable. Both overuse and underuse reflect the fundamental difficulty of matching people and treatments in the face of great heterogeneity and uncertain diagnosis.

Moderate and serious mental illnesses, even when appropriately treated, generate significant costs for individuals and impose important externalities on others. The public goods problems associated with addressing these adverse risks and externalities generate underinvestment in complementary social supports and human services like supported housing, supported employment, income supports, and care management. These complementary services can directly reduce the individual and societal consequences of illness and likely also improve the efficacy of clinical treatment.

Economic research and evidence can improve the design of health insurance policies and the level and rationing systems for social and income supports. The task, however, will be challenging. Issues of stigma and our incomplete ability to distinguish between levels of need and to match conditions and treatments will continue to test our abilities to cope with mental illnesses efficiently and humanely.

■ *We gratefully acknowledge financial support from the Robert Wood Johnson Foundation. Caitlin Rowley, Karina Aguilar, and Dong Ding provided expert research assistance and editorial support.*

Referencßes

ADA National Network. "What Is the Americans with Disabilities Act (ADA)?" https://adata.org/learn-about-ada/ (accessed August 17, 2020).

Adler, David A., Thomas J. McLaughlin, William H. Rogers, Hong Chang, Leueen Lapitsky, and Debra Lerner. 2006. "Job Performance Deficits Due to Depression." *American Journal of Psychiatry* 163 (9): 1569–76.

Anesetti-Rothermel, Andrew, and Usha Sambamoorthi. 2011. "Physical and Mental Illness Burden: Disability Days among Working Adults." *Population Health Management* 14 (5): 223–30.

Areán, Patricia A., and Anna Ratzliff. 2020. "Balancing Access to Medications and Psychosocial Treatments." In *The Palgrave Handbook of American Mental Health Policy*, edited by Howard H. Goldman, Richard G. Frank, and Joseph P. Morrissey, 101–28. Cham, Switzerland: Palgrave Macmillan.

Bagalman, Erin, and Ada S. Cornell. 2018. *Prevalence of Mental Illness in the United States: Data Sources and Estimates*. Washington, DC: Congressional Research Service. https://sgp.fas.org/crs/misc/R43047.pdf.

Ban, Thomas A. 2007. "Fifty Years Chlorpromazine: A Historical Perspective." *Neuropsychiatric Disease and Treatment* 3 (4): 495–500.

Bee, Penny E., Peter Bower, Simon Gilbody, and Karina Lovell. 2010. "Improving Health and Productivity of Depressed Workers: A Pilot Randomized Controlled Trial of Telephone Cognitive Behavioral Therapy Delivery in Workplace Settings." *General Hospital Psychiatry* 32 (3): 337–40.

Berndt, Ernst R., Lorrin M. Koran, Stan N. Finkelstein, Alan J. Gelenberg, Susan G. Kornstein, Ivan M. Miller, Michael E. Thase, George A. Trapp, and Martin B. Keller. 2000. "Lost Human Capital from Early-Onset Chronic Depression." *American Journal of Psychiatry* 157 (6): 940–47.

Blier, Pierre, Martin B. Keller, Mark H. Pollack, Michael E. Thase, John M. Zajecka, and David L. Dunner. 2007. "Preventing Recurrent Depression: Long-Term Treatment for Major Depressive Disorder." *Journal of Clinical Psychiatry* 68 (3): e06.

Bond, Gary R., Sunny Jung Kim, Deborah R. Becker, Sarah J. Swanson, Robert E. Drake, Izabela M. Krzos, Virginia V. Fraser, Sheila O'Neill, and Rochelle L. Frounfelker. 2015. "A Controlled Trial of Supported Employment for People with Severe Mental Illness and Justice Involvement." *Psychiatric Services* 66 (10): 1027–34.

Brand, Jennie E., Becca R. Levy, and William T. Gallo. 2008. "Effects of Layoffs and Plant Closings on Subsequent Depression among Older Workers." *Research on Aging* 30 (6): 701–21.

Bronson, Jennifer, and Marcus Berzofsky. 2017. "Indicators of Mental Health Problems Reported by Prisoners and Jail Inmates, 2011-12." *Bureau of Justice Statistics Special Report.* Washington, DC: Bureau of Justice Statistics.

Bureau of Labor Statistics. 2022. *Occupational Outlook Handbook.* US Department of Labor. https://www.bls.gov/ooh/. (accessed December, 6 2022).

Carson, E. Ann. 2021. *Prisoners in 2020—Statistical Tables.* Washington, DC: Bureau of Justice Statistics.

Chernew, Michael E., and Joseph P. Newhouse. 2011. "Health Care Spending Growth." In *Handbook of Health Economics*, Vol. 2, edited by Mark V. Pauly, Thomas G. Mcguire, and Pedro P. Barros, 1–43. Amsterdam: North-Holland.

Christensen, Maria K., John J. McGrath, Natalie C. Momen, Harvey A. Whiteford, Nanna Weye, Esben Agerbo, Casten Bocker Pedersen, Preben Bo Mortensen, Oleguer Plana-Ripoll, and Kim Moesgaard Iburg. 2022. "The Cost of Mental Disorders in Denmark: A Register-Based Study." *npj Mental Health Research* 1: 1–7.

Copeland, William E., Guangyu Tong, Lauren Gaydosh, Sherika Hill, Jennifer Godwin, Lilly Shanahan, and Elizabeth Costello. 2022. "Twenty Year Outcomes of an Unconditional Cash Transfer." https://doi.org/10.2139/ssrn.4059228.

Costello, E. Jane, Scott N. Compton, Gordon Keeler, and Adrian Angold. 2003. "Relationships between Poverty and Psychopathology: A Natural Experiment." *JAMA* 290 (15): 2023–9.

Cuddeback, Gary S., Jennie M. Simpson, and Juliet C. Wu. 2020. "A Comprehensive Literature Review of Forensic Assertive Community Treatment (FACT): Directions for Practice, Policy and Research." *International Journal of Mental Health* 49 (2): 106–27.

Cullen, Kimberly L., Emma Irvin, Alex Collie, Fiona Clay, Ulrik Gensby, Paul A. Jennings, Sheilah Hogg-Johnson, et al. 2017. "Effectiveness of Workplace Interventions in Return-to-Work for Musculoskeletal, Pain-Related and Mental Health Conditions: An Update of the Evidence and Messages for Practitioners." *Journal of Occupational Rehabilitation* 28 (1): 1–15.

Danziger, Sheldon, Richard G. Frank, and Ellen Meara. 2009. "Mental Illness, Work, and Income Support Programs." *American Journal of Psychiatry* 166 (4): 398–404.

Dieleman, Joseph L., Jackie Cao, Abby Chapin, Carina Chen, Zhiyin Li, Angela Liu, Cody Horst, et al. 2020. "US Health Care Spending by Payer and Health Condition, 1999–2016." *JAMA* 323 (9): 863–84.

Dow, William H., Anna Godøy, Christopher Lowenstein, and Michael Reich. 2020. "Can Labor Market Policies Reduce Deaths of Despair?" *Journal of Health Economics* 74: 102372.

Drukarch, Benjamin, Gabriel E. Jacobs, & Micha M. M. Wilhelmus. 2020. "Solving the Crisis in Psychopharmacological Research: Cellular-Membrane(s) Pharmacology to the Rescue?" *Biomedicine and Pharmacotherapy* 130: 110545.

Evans, William N., and Craig L. Garthwaite. 2014. "Giving Mom a Break: The Impact of Higher EITC Payments on Maternal Health." *American Economic Journal: Economic Policy* 6 (2): 258–90.

Fletcher, Jason. 2013. "Adolescent Depression and Adult Labor Market Outcomes." *Southern Economic Journal* 80 (1): 26–49.

Frank, Richard G., and Sherry A. Glied. 2006. *Better but Not Well: Mental Health Policy in the United States since 1950.* Baltimore: Johns Hopkins University Press.

Frank, Richard G., and Sherry A. Glied. 2023. "Replication data for: America's Continuing Struggle with Mental Illnesses: Economic Considerations." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E183993V1.

Frank, Richard G., Kirsten Beronio, and Sherry A. Glied. 2014. "Behavioral Health Parity and the Affordable Care Act." *Journal of Social Work in Disability and Rehabilitation* 13 (1–2): 31–43.

Germack, Hayley D., Coleman Drake, Julie M. Donohue, Ezra Golberstein, and Susan H. Busch. 2020. "National Trends in Outpatient Mental Health Service Use among Adults between 2008 and 2015." *Psychiatric Services* 71 (11): 1127–35.

Glied, Sherry A., Richard G. Frank, and Joanna Wexler. 2020. "Mental Health Disability, Employment, and Income Support in the Twenty-First Century." In *The Palgrave Handbook of American Mental Health Policy*, edited by Howard H. Goldman, Richard G. Frank, and Joseph P. Morrissey, 659–77. Cham, Switzerland: Palgrave Macmillan.

Goldman, Howard H., Richard G. Frank, and Joseph P. Morrissey, eds. 2020. *The Palgrave Handbook of American Mental Health Policy*. Cham, Switzerland: Palgrave Macmillan.

Granger, Barbara, Richard Baron, and Susan Robinson. 1997. "Findings from a National Survey of Job Coaches and Job Developers about Job Accommodations Arranged between Employers and People with Psychiatric Disabilities." *Journal of Vocational Rehabilitation* 9 (3), 235–51.

Grob, Gerald N. 1994. *The Mad Among Us: A History of the Care of America's Mentally Ill.* New York: Free Press.

Hakulinen, C., M. Elovainio, M. Arffman, S. Lumme, S. Pirkola, I. Keskimäki, K. Manderbacka, and P. Böckerman. 2019. "Mental Disorders and Long-Term Labour Market Outcomes: Nationwide Cohort Study of 2 055 720 Individuals." *Acta Psychiatrica Scandinavica* 140 (4): 371–81.

Hickox, Stacy A., and Angela Hall. 2018. "Atypical Accommodations for Employees with Psychiatric Disabilities." *American Business Law Journal* 55 (3): 537–94.

Horowitz, Jake. 2021. "Local Spending on Jails Tops $25 Billion in Latest Nationwide Data." Pew Charitable Trusts, https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2021/01/local-spending-on-jails-tops-$25-billion-in-latest-nationwide-data.

Horvitz-Lennon, Marcela. 2020. "Evidence-Based Practices or Practice-Based Evidence: What Is the Future?" In *The Palgrave Handbook of American Mental Health Policy,* edited by Howard H. Goldman, Richard G. Frank, and Joseph P. Morrissey, 603–38. Cham, Switzerland: Palgrave Macmillan.

Insel, Thomas R. 2022. *Healing: Our Path from Mental Illness to Mental Health.* New York: Penguin Press.

Kane, John M., Delbert G. Robinson, Nina R. Schooler, Kim T. Mueser, David L. Penn, Robert A. Rosenheck, Jean Addington, et al. 2016. "Comprehensive versus Usual Community Care for First-Episode Psychosis: 2-Year Outcomes from the NIMH RAISE Early Treatment Program." *American Journal of Psychiatry* 173 (4): 362–72.

Keisler-Starkey, Katherine, and Lisa N. Bunch. 2021 *Health Insurance Coverage in the United States: 2020.* Washington, DC: US Government Publishing Office.

Kessler, Ronald C. 1994. "The National Comorbidity Survey of the United States." *International Review of Psychiatry* 6 (4): 365–76.

Kessler, Ronald C., Ellen E. Walters, and Melinda S. Forthofer. 1998. "The Social Consequences of Psychiatric Disorders, III: Probability of Marital Stability." *American Journal of Psychiatry* 155 (8): 1092–6.

Kröger, Christoph, Katharina Bode, Eva-Maria Wunsch, Sören Kliem, Anja Grocholewski, and Friederike Finger. 2015. "Work-Related Treatment for Major Depressive Disorder and Incapacity to Work: Preliminary Findings of a Controlled, Matched Study." *Journal of Occupational Health Psychology* 20 (2): 248–58.

Lagerveld, Suzanne E., Roland W.B. Blonk, Veerle Brenninkmeijer, Leoniek Wijngaards-de Meij, and Wilmar B. Schaufeli. 2012. "Work-Focused Treatment of Common Mental Disorders and Return to Work: A Comparative Outcome Study." *Journal of Occupational Health Psychology* 17 (2): 220–34.

Latimer, Eric A. 1999. "Economic Impacts of Assertive Community Treatment: A Review of the Literature." *Canadian Journal of Psychiatry* 44 (5): 443–54.

Leopold, Josh. 2020. "Housing for People with Serious Mental Illness." In *The Palgrave Handbook of American Mental Health Policy*, edited by Howard H. Goldman, Richard G. Frank, and Joseph P. Morrissey, 389–408. Cham, Switzerland: Palgrave Macmillan.

Lerner, Debra, David A. Adler, Hong Chang, Leueen Lapitsky, Maggie Y. Hood, Carla Perissinotto, John Reed, Thomas J. McLaughlin, Ernst R. Berndt, and William H. Rogers. 2004. "Unemployment, Job Retention, and Productivity Loss among Employees with Depression." *Psychiatric Services* 55 (12): 1371–8.

Lerner, Debra, David Adler, Richard C. Hermann, Hong Chang, Evette J. Ludman, Annabel Greenhill, Katherine Perch, William C. McPeck, and William H. Rogers. 2012. "Impact of a Work-Focused Intervention on the Productivity and Symptoms of Employees with Depression." *Journal of Occupational and Environmental Medicine* 54 (2): 128–35.

Lerner, Debra, David A. Adler, William H. Rogers, Hong Chang, Annabel Greenhill, Elina Cymerman, and Francisca Azocar. 2015. "A Randomized Clinical Trial of a Telephone Depression Intervention to Reduce Employee Presenteeism and Absenteeism." *Psychiatric Services* 66 (6): 570–77.

Lerner, Debra, David A. Adler, William H. Rogers, Hong Chang, Annabel Greenhill, and Francisca Azocar. 2017. "The Double Burden of Work Stress and Depression: A Workplace Intervention." In *The Handbook of Stress and Health: A Guide to Research and Practice*, edited by Cary L. Cooper and James Campbell Quick, 147–67.

Levine, Daniel S., and Dianne R. Levine. 1975. *The Cost of Mental Illness: 1971.* Washington, DC; US Government Printing Office.

Lucas, David S. 2017. "The Impact of Federal Homelessness Funding on Homelessness." *Southern Economic Journal* 84 (2): 548–76.

Ly, Angela, and Eric Latimer. 2015. "Housing First Impact on Costs and Associated Cost Offsets: A Review of the Literature." *Canadian Journal of Psychiatry* 60 (11): 475–87.

Maruschak, Laura, and Jennifer Bronson. 2021. *Survey of Prison Inmates, 2016: Indicators of Mental Health Problems Reported by Prisoners.* Washington, DC: Bureau of Justice Statistics.

Mechanic, David, Donna D. McAlpine, and David A. Rochefort. 2014. *Mental Health and Social Policy: Beyond Managed Care.* 6th ed. Boston, MA: Pearson.

Millan, Mark J., Yves Agid, Martin Brüne, Edward T. Bullmore, Cameron S. Carter, Nicola S. Clayton, Richard Connor, et al. 2012. "Cognitive Dysfunction in Psychiatric Disorders: Characteristics, Causes and the Quest for Improved Therapy." *Nature Reviews Drug Discovery* 11 (2): 141–68.

Mintz, Jim, Lois Imber Mintz, Mary Jane Arruda, and Sun Sook Hwang. 1992. "Treatments of Depression and the Functional Capacity to Work." *Archives of General Psychiatry* 49 (10): 761–68.

Newman, Sandra J., James D. Reschovsky, Keith Kaneda, and Anne M. Hendrick. 1994. "The Effects of Independent Living on Persons with Chronic Mental Illness: An Assessment of the Section 8 Certificate Program." *Milbank Quarterly* 72 (1): 171–98.

Office of Technology Assessment. 1980. "Background Paper #3: The Efficacy and Cost Effectiveness of Psychotherapy." *Cost-Effectiveness of Medical Technology*. Washington, DC: US Government Printing Office. https://www.princeton.edu/~ota/disk3/1980/8020/8020.PDF.

O'Flaherty, Brendan. 1998. *Making Room: The Economics of Homelessness.* Cambridge, MA: Harvard University Press.

Osher, Fred, and Michael Thompson. 2020. "Adults with Serious Mental Illnesses Who Are Arrested and Incarcerated." In *The Palgrave Handbook of American Mental Health Policy*, edited by Howard H. Goldman, Richard G. Frank, and Joseph P. Morrissey, 471–508. Cham, Switzerland: Palgrave Macmillan.

Owens, Pamela L., Kathryn R. Fingar, Kimberly W. McDermott, Pradip K. Muhuri, and Kevin C. Heslin. 2019. "Inpatient Stays Involving Mental and Substance Use Disorders, 2016" Agency for Healthcare Research and Quality Statistical Brief #249.

Productivity Commission of the Australian Government. 2020. *Mental Health: Productivity Commission Inquiry Report*, Vol. 1. Canberra: Australian Government.

Raphael, Steven, and Michael A. Stoll. 2013. "Assessing the Contribution of the Deinstitutionalization of the Mentally Ill to Growth in the US Incarceration Rate." *The Journal of Legal Studies* 42 (1): 187–222.

Reeves, Aaron, Martin McKee, Johan Mackenbach, Margaret Whitehead, and David Stuckler. 2017. "Introduction of a National Minimum Wage Reduced Depressive Symptoms in Low-Wage Workers: A Quasi-natural Experiment in the UK." *Health Economics*, 26 (5): 639–55.

Regier, Darrel A., William E. Narrow, Donald S. Rae, Ronald W. Manderscheid, Ben Z. Locke, and Frederick K. Goodwin. 1993. "The De Facto US Mental and Addictive Disorders Service System." *Archives of General Psychiatry* 50 (2): 85–94.

Rock, Philippa L., Jonathan P. Roiser, Wim J. Riedel, and Andrew D. Blackwell. 2013. "Cognitive Impairment in Depression: A Systematic Review and Meta-analysis." *Psychological Medicine* 44 (10): 2029–40.

Santo L. and Okeyode T. 2018. "National Ambulatory Medical Care Survey: 2018 National Summary Tables." Hyattsville, MD: National Center on Health Statistics.

Saunders, Heather, and Robin Rudowitz. 2022. "Demographics and Health Insurance Coverage on Non-Elderly Adults with Mental Illness and Substance Use Disorders in 2020." Kaiser Family Foundation, June 6. https://www.kff.org/medicaid/issue-brief/demographics-and-health-insurance-coverage-of-nonelderly-adults-with-mental-illness-and-substance-use-disorders-in-2020/.

Scott, Jack E., and Lisa B. Dixon. 1995. "Assertive Community Treatment and Case Management for Schizophrenia." *Schizophrenia Bulletin* 21 (4): 657–68.

Sim, Kang, Wai Keat Lau, Jordan Sim, Min Yi Sum, and Ross J. Baldessarini. 2015. "Prevention of Relapse and Recurrence in Adults with Major Depressive Disorder: Systematic Review and Meta-Analyses of Controlled Trials." *International Journal of Neuropsychopharmacology* 19 (2): pyv076.

Steadman, Henry J., Lisa Callahan, Pamela Clark Robbins, Roumen Vesselinov, Thomas G. McGuire, and Joseph P. Morrissey. 2014. "Criminal Justice and Behavioral Health Care Costs of Mental Health Court Participants: A Six-Year Study. *Psychiatric Services* 65 (9): 1100–4.

Stein, Leonard I., and Alberto B. Santos. 1998. *Assertive Community Treatment of Persons with Severe Mental Illness.* New York: W. W. Norton & Company.

Substance Abuse and Mental Health Services Administration (SAMHSA). 2014. "Projections of National Expenditures for Treatment of Mental and Substance Use Disorders, 2010–2020." SAMHSA Publications and Digital Products, HHS Publication No. SMA-14-4883.

Substance Abuse and Mental Health Services Administration (**SAMHSA**). 2016. *Behavioral Health Spending and Use Accounts, 1986–2014.* Rockville, MD: Substance Abuse and Mental Health Services Administration.

Substance Abuse and Mental Health Services Administration (SAMHSA). 2021. "2020 National Survey of Drug Use and Health (NSDUH) Detailed Tables." Rockville, MD: Substance Abuse and Mental Health Services Administration.

Swanson, Jeffrey W., Colleen L. Barry, and Marvin S. Swartz. 2020. "Gun Violence Prevention and Mental Health Policy." In *The Palgrave Handbook of American Mental Health Policy*, edited by Howard H. Goldman, Richard G. Frank, and Joseph P. Morrissey, 509–41. Cham, Switzerland: Palgrave Macmillan.

Terlizzi, Emily P, and Benjamin Zablotsky. 2020. "Mental Health Treatment among Adults: United States, 2019." National Center for Health Statistics (NCHS) Data Brief No. 380.

Treatment Advocacy Center. 2019. "Road Runners: The Role and Impact of Law Enforcement in Transporting Individuals with Severe Mental Illness." Treatment Advocacy Center, May. https://www.treatmentadvocacycenter.org/road-runners.

Tsai, Daniel. 2021. "Medicaid Guidance on the Scope of and Payments for Qualifying Community-Based Mobile Crisis Intervention Services." Baltimore: US Department of Health and Human Services, Centers for Medicare & Medicaid. https://www.medicaid.gov/federal-policy-guidance/downloads/sho21008.pdf.

Tyrer, Peter J. 2009. "Twisted Science, Regulation, and Molecules." *The Lancet* 373 (9674): 1513–14.

Urban Institute. 2020. "Criminal Justice Expenditures: Police, Corrections, and Courts." https://www.urban.org/policy-centers/cross-center-initiatives/state-and-local-finance-initiative/state-and-local-backgrounders/criminal-justice-police-corrections-courts-expenditures.

US Department of Agriculture (USDA). 2020. Supplemental Nutrition Assistance Program: SNAP Web Tables. https://www.fns.usda.gov/pd/supplemental-nutrition-assistance-program-snap (accessed June 22, 2022).

US Department of Health and Human Services (HHS). 2020. "Characteristics and Financial Circumstances of TANF Recipients, Fiscal Year 2019." Office of Family Assistance, November 5. https://www.acf.hhs.gov/ofa/data/characteristics-and-financial-circumstances-tanf-recipients-fiscal-year-2019.

US Department of Housing and Urban Development (HUD). 2018a. "Community Planning and Development Program Formula Allocations for FY 2018." https://www.hud.gov/program_offices/comm_planning/budget/fy18/ (accessed June 22, 2022).

US Department of Housing and Urban Development (HUD). 2018b. "Fiscal Year 2018 Continuum of Care Competition Homeless Assistance Award Report." https://www.hudexchange.info/sites/onecpd/assets/File/2018-all-coc-grants.pdf (accessed June 22, 2022).

US Department of Housing and Urban Development (HUD). 2019. "2019 Continuum of Care Homeless Assistance Programs Homeless Populations and Subpopulations." https://files.hudexchange.info/reports/published/CoC_PopSub_NatlTerrDC_2019.pdf (accessed June 22, 2022).

US Department of Housing and Urban Development (HUD). 2020. "Assisted Housing: National and Local." https://www.huduser.gov/portal/datasets/assthsg.html (accessed June 22, 2022).

US Department of Justice, Bureau of Prisons (BOP). 2021. "Annual Determination of Average Cost of

Incarceration Fee (COIF)." US Department of Justice, September 1. https://www.federalregister.gov/documents/2021/09/01/2021-18800/annual-determination-of-average-cost-of-incarceration-fee-coif/.

US Department of Justice, National Institute of Correction (NIC). 2019. "2019 National Averages." https://nicic.gov/state-statistics/2019 (accessed May 11, 2022.

US Social Security Administration (SSA). 2020a. "Annual Statistical Report on the Social Security Disability Insurance Program, 2019." US Social Security Administration, October. https://www.ssa.gov/policy/docs/statcomps/di_asr/index.html.

US Social Security Administratio (SSA). 2020b. "SSI Annual Statistical Report, 2019." US Social Security Administration, August. https://www.ssa.gov/policy/docs/statcomps/ssi_asr/2019.

US Social Security Administration (SSA). 2022. "Ticket to Work." https://choosework.ssa.gov/about/how-it-works/index.html (accessed July 27, 2022).

US Supreme Court. 1999. Olmstead v. L. C.

Wang, Buyi, Richard G. Frank, and Sherry A. Glied. 2022. "Lasting Scars: The Impact of Depression in Early Adulthood on Subsequent Labor Market Outcomes." National Bureau of Economic Research Working Paper 30776.

Zeng, Zhen, and Todd Minton. 2021. *Jail Inmates in 2019*. Washington, DC: Bureau of Justice Statistics.

Zwerling, Craig, Paul S. Whitten, Nancy L. Sprince, Charles S. Davis, Robert B. Wallace, Peter Blanck, and Steven G. Heeringa. 2003. "Workplace Accommodations for People with Disabilities: National Health Interview Survey Disability Supplement, 1994–1995." *Journal of Occupational and Environmental Medicine* 45 (5): 517–25.

Journal of Economic Perspectives—Volume 37, Number 2—Spring 2023—Pages 179–202

Depression and Loneliness among the Elderly in Low- and Middle-Income Countries

Abhijit Banerjee, Esther Duflo, Erin Grela, Madeline McKelway, Frank Schilbach, Garima Sharma, and Girija Vaidyanathan

The elderly population is growing rapidly in low- and middle-income countries—it is projected to increase from about 500 million in 2019 to over 1.2 billion in 2050 (UN DESA Population Division 2019)—yet the well-being and mental health of this population are not a policy or research priority.

This situation has arisen for two main reasons. First, issues facing older people are not a general policy or research priority in low- and middle-income countries, perhaps because populations tend to skew younger. For instance, of the 528 studies used by the Global Burden of Disease, only 17 (covering just six countries) were designed to study the elderly in low- and middle-income countries. Second, mental health issues are generally underemphasized in low- and middle-income countries. High-income countries allocate about 3.4 percent of their total government health expenditure to mental health, compared to 0.3 percent in low- and middle-income countries—and only 0.09 percent in the nine low-income countries covered by the WHO Mental Health Atlas (Ridley et al. 2020). This lack of spending results in very limited availability of trained staff and treatment: there are 1.4 mental health workers per 100,000 population in poor countries, compared to 62 per 100,000 in

■ *Abhijit Banerjee is Ford Foundation International Professor of Economics, Esther Duflo is Abdul Latif Jameel Professor of Poverty Alleviation and Development Economics, Erin Grela is a PhD student in economics, Frank Schilbach is Gary Loveman Career Development Associate Professor of Economics, and Garima Sharma is a PhD student in economics, all at the Massachusetts Institute of Technology, Cambridge, Massachusetts. Madeline McKelway is Assistant Professor of Economics, Dartmouth College, Hanover, New Hampshire. Girija Vaidyanathan is Professor of Practice in the Humanities and Social Sciences Department, Indian Institute of Technology, Madras, India.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.179.

rich countries, and over 90 percent of cases of major depressive disorder are untreated in low- and middle-income countries (Thornicroft et al. 2017).

Similarly, research and data are sparse: less than 2.7 percent of published research on public health focuses on mental health in low- and middle-income countries, compared to 8 percent in rich countries (World Health Organization 2021). In the Global Burden of Disease database, no data on mental health exists for 88 of the 134 low- and middle-income countries. For the countries with data, the sample sizes of the underlying studies are typically small, with measurements of mental health coming from short screening instruments.

In this essay, we begin by shining a spotlight on an unseen epidemic of poor mental health among the elderly in developing countries. We use a set of existing, high-quality surveys with well-validated survey tools for measuring depression. We create comparable estimates of the prevalence of depression among people aged 55 and up across seven low- and middle-income countries and compare those to the United States.

Our first key finding is that the prevalence of symptoms of depression among the elderly is much higher in poorer countries than it is in the United States. For both men and women, in every age range, the rates of depression symptoms are lower in the United States than nearly all our comparison countries. For example, in India, 26 percent of men and 31 percent of women aged 61–70 have symptoms indicating high likelihood of depression, compared to 11 percent of men and 14 percent of women aged 61–70 in the United States.

Second, many of the elderly in low-income countries feel lonely, despite the common presumption that most elderly in these countries live with their family. In several of the countries in our data sets, the fraction of elderly who report feeling lonely is largely in line with that of the United States: between 10 and 25 percent, varying slightly with age. In other countries, the loneliness rate is much higher than in the United States, reaching over 50 percent in Mexico among those over 80 years old. Importantly, loneliness is a strong predictor of depression.

Our third key finding is that poor mental health is also associated with poor physical health and an elevated risk of death within the next two years. If these associations are causal, it would imply that treating and preventing mental health diseases could be an important policy instrument to facilitate healthy aging more broadly. The effects would be especially substantial in countries where the share of the older population is poised to increase substantially in the next few decades.

Prior research highlights a few key factors associated with poor mental health among older people. To explore these patterns in more detail in one middle-income country, we then turn to a new panel survey on the mental health of the elderly we conducted in Tamil Nadu, India, where we collected detailed measurements on the correlates of depression. One advantage of the survey is that it deliberately oversamples the elderly living alone, an often-ignored group that we show to be especially at risk for depression and functional limitations. We show that physical health challenges, poverty, and social isolation (as measured by living alone) are strongly correlated with depression.

Finally, we draw on these findings, results from our field experiments in Tamil Nadu, and the existing literature on mental health among the elderly and in the general population to propose some promising policy interventions to address elderly mental health in poor countries. We suggest some priorities for future research and policy action on the topic.

Measuring Mental Health in Low- and Middle-Income Countries

Creating Internationally Comparable Mental Health Measures

To construct robust and comparable data on mental health of the elderly in low- and middle-income countries, we combine data from publicly available surveys of six countries: Brazil (Lima-Costa et al. 2018), China (Zhao et al. 2014), Costa Rica (Rosero-Bixby, Fernández, and Dow), India (Bloom, Sekher, and Lee 2021), Mexico (Wong, Michaels-Obregon, and Palloni 2017), and South Africa (Berkman 2023).[1] These surveys are part of a family of surveys modeled after the US Health and Retirement Study (Bugliari et al. 2022), an ongoing, decades-long longitudinal study that has become the template for a growing network of longitudinal aging studies around the world. With support from the National Institute of Aging in the United States, the different research organizations that collected these data have made explicit efforts to harmonize survey instruments and collection procedures, all part of the "Health and Retirement Family of Surveys."[2] We use this data to make cross-country comparisons, while including data from the 2014 and 2016 wave of this study in the United States as an additional benchmark. To have more data for Africa, we also include data from one other survey that followed a similar template and asked detailed questions pertaining to mental health: the Malawi Longitudinal Study of Families and Health (Kohler et al. 2013).

The Diagnostic and Statistical Manual of Mental Disorders describes depression as a family of disorders characterized by "the presence of sad, empty, or irritable mood, accompanied by related changes that significantly affect the individual's capacity to function" (American Psychiatric Association 2022). Within this family, Major Depressive Disorder is often referred to as "clinical depression." Major Depressive Disorder consists of nine symptoms; to be diagnosed, one must show at least five of those symptoms within the past two weeks. Of those five, at least one must be "a depressed mood" or "loss of interest or pleasure in all, or almost all, activities" for most of the day, nearly every day.

Measuring depression prevalence among the elderly is particularly challenging, as older adults who experience depression tend to show different symptoms than

[1] The studies are documented by the Program on Global Aging, Health, and Policy at the University of Southern California and instructions for accessing the data can be found at https://g2aging.org/survey-overview.

[2] In the online Appendix, we provide detailed information on how the data is constructed to ensure comparability of the measures in the different datasets.

younger adults. For example, older adults may report a lack of emotions rather than a depressed mood (Blazer and Hybels 2014). In light of this, surveys of older adults often use adapted versions of questions about symptoms to screen for depression, which we discuss in more detail below.

The gold standard for diagnosis of depression is a face-to-face interview with a trained psychiatrist, but this is not feasible at scale in low-income settings. Instead, the mental health portion of the questionnaires in the Health and Retirement Family of Surveys includes between 8 and 15 questions about recent experience of depression symptoms, such as, "How often during the past week did you enjoy life?" or "How often during the past week did you feel that everything you did was an effort?" This reflects a compromise between a lengthy interview, which may elicit low response rates or low response quality, and the very short screening surveys used by the Gallup Polls or the World Health Organization, which may be a bit too coarse to reliably capture the true prevalence of depression. The questionnaires used in our analysis have been validated for studying depression among the elderly in studies that compared their prediction to an evaluation performed by a trained psychiatrist (Vilagut et al. 2016).

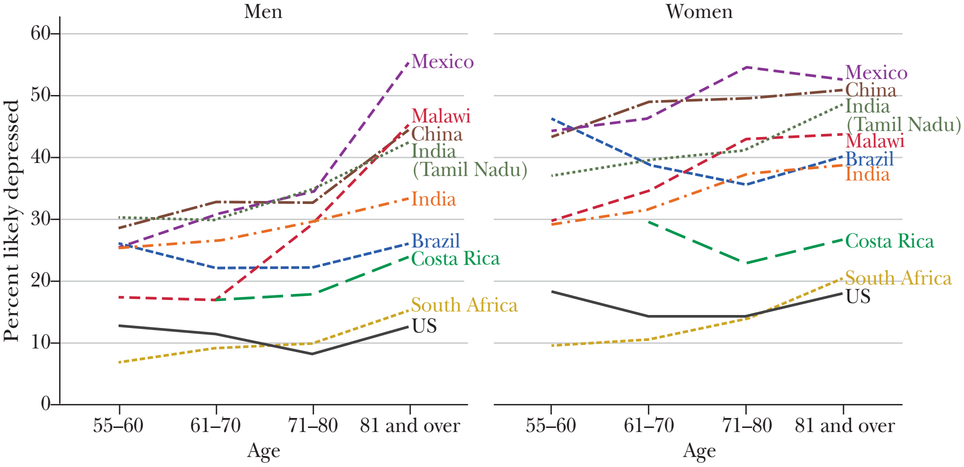
A difficulty that arises in this kind of work, especially in low- and middle-income countries where mental health is not commonly discussed, is that mental health questionnaires are by nature sensitive to issues of translation, interpretation, and cultural norms. Thus, comparing depression levels across countries, even when using the same (translated and validated) questionnaire, can be difficult. An individual is considered to likely be depressed if they report a number of depression symptoms above the questionnaire's cutoff point. To select a cutoff point, validation studies balance the sensitivity (the rate of false negatives) and the specificity (the rate of false positives) of the measurement, but they often find different optimal cutoff points in different settings. To compare depression prevalence across countries, we select cutoff scores for each country based on existing literature, with the aim of identifying people who, in each context, would be described as likely depressed.

We acknowledge that the choice of medium-length survey instrument for measuring depression in these datasets is driven by what is feasible to use on large sample sizes rather than a belief that these measures are without flaws. Compared to clinical diagnosis or even to lengthier survey instruments, these surveys are prone to measurement error and may lead to an overestimate of clinical depression. This is particularly true when using an imperfect screening test for a condition that has low prevalence in the population—and therefore produces many more chances to produce false positives than false negatives (Maxim, Niebo, and Utell 2014). Conversely, persons who are reluctant to introspect may not engage with the survey, which would lead to false negatives.

Fortunately, these classification errors are perhaps not critical for the purpose of this essay. In our analysis, we report the fraction of individuals who are "likely to be depressed" because it is an easily interpretable, comparable summary measure. However, regardless of whether a tag of being "depressed" translates precisely into formal clinical depression or not, there is no doubt that depression *symptoms* are associated with low well-being.

Figure 1

Prevalence of Depression Symptoms by Gender, Age, and Country



Source: The data for the United States uses sample averages from the Health and Retirement Study in 2014. The data for Malawi comes from another independent study (Malawi Longitudinal Study of Families and Health) of the health and well-being of older adults. Data for Tamil Nadu comes from our own study that was conducted in the state of Tamil Nadu. The data for the other countries comes from nationally representative studies modeled after the US Health and Retirement Survey (Brazil, China, Costa Rica, Mexico, India, and South Africa). When available, survey weights are used to calculate averages that are nationally representative of older adults in each country or area.

Notes: This figure shows the share of the population that is likely to be depressed (as determined by whether the depression index scores surpass the specified thresholds) for each age group and gender across the countries in our sample.

Prevalence of Depression among the Elderly

Figure 1 shows the fraction of the elderly who are likely to be depressed, as determined by whether the number of symptoms reported on a standard depression screening questionnaire falls above the country-specific threshold.[3] The figure illustrates two important patterns.

First, the prevalence of depression symptoms among the elderly in poorer countries is typically high relative to the United States. For example, for all low- and middle-income countries in our dataset except for South Africa, about 20–35 percent of men between the ages of 71 and 80 show symptoms indicative of depression—more than double the US rates. China and Mexico have the highest prevalence of depressive symptoms, and Costa Rica and South Africa have the lowest. India is in the middle. For almost every combination of country and age category, depression rates are higher for women than for men, consistent with higher prevalence

[3] The supplementary Appendix includes additional figures and tables with further empirical results. Throughout the remainder of our text, we mention and discuss these results without explicitly referring to the corresponding Appendix figures and tables.

of anxiety and depression among women in many other contexts (Salk, Hyde, and Abramson 2017).

Second, the prevalence of depression symptoms increases with age in low- and middle-income countries, with particularly pronounced depressive symptoms at age 70 and above. This result appears to contrast with previous work, mostly in high-income countries, that documents a U-shaped pattern of well-being (and/or an inverse U-shape for mental distress) in age (for example, Blanchflower 2021; Giuntella et al. 2023). In line with this literature, we find depression does not increase with age in the United States.[4] We can only speculate as to why we see depression rise with age in low- and middle-income countries but not in the United States. Perhaps greater access to healthcare in the United States allows individuals to age with relatively fewer health issues and pain. These patterns could also reflect greater ability to save for retirement in the United States, or pensions that are more generous or widespread relative to our comparison countries.

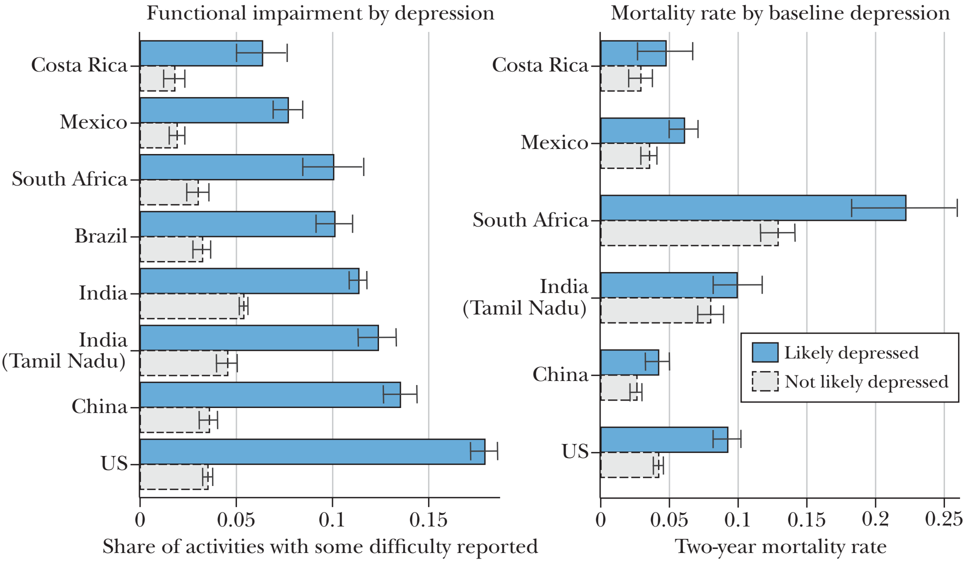
The data also suggest that elderly depression remains largely undiagnosed and untreated in low-income settings, consistent with enormous treatment gaps for depression that have been documented for the general population (Thornicroft et al. 2017). In China, for example, among the approximately 35 percent of respondents who show signs of depression, only 2 percent have ever been diagnosed by a doctor with any psychological condition, 1 percent are taking any medication for psychological conditions, and 0.3 percent are receiving any other type of treatment.

Depression, Functional Abilities, and Death

Depression is a key predictor and aspect of poor well-being and low life satisfaction (Kahneman and Krueger 2006). In addition, it is associated with a host of other adverse outcomes. We focus on the relationship between depression and functional impairment because aging is associated with a slow deterioration in the ability to carry out basic "activities of daily living," which can lead to a considerable loss in life quality and autonomy. The surveys in our analysis typically cover six activities of daily living: dressing, eating, bathing, getting out of bed, using the toilet, and controlling urination and defecation. Respondents are asked whether they can do a particular activity easily, with difficulty, or not at all. To construct a summary

[4] Much of the psychiatric literature, based on evidence mostly from rich countries, finds depression decreasing into old age (Eaton et al. 1989), but this is the topic of an active debate (Yang 2007). In the economics literature, Giuntella el al. (2023) examine depression, anxiety, and other measures of well-being in rich countries and find a "midlife crisis" around age 50. Blanchflower (2021) comes to a similar conclusion using datasets that cover the majority of countries in the world. The latter findings are not necessarily inconsistent with our results for three reasons. First, Blanchflower primarily focuses on people aged 70 or younger, whereas we include all those over the age of 55. Thus, it is possible that well-being declines in middle age, improves in late adulthood, and then declines again in the oldest age ranges. Second, our data features validated surveys specifically designed for the elderly as opposed to single-item surveys asking about overall well-being or *lifetime* prevalence of depression. Finally, Blanchflower's results are especially strong upon inclusion of controls such as job status and marital status; we omit these controls because we contend that widowhood and unemployment are integral aspects of aging, and a channel through which mental well-being may decline.

Figure 2
Correlates of Depression: Functional Impairment and Mortality by Country



Source: Same as Figure 1. In addition, the mortality figure uses one additional wave of the data than what is used in Figure 1. For each country, we use information on respondent status in the subsequent wave. The years and details for each country are listed in Appendix Table 1.
Note: The left panel of the figure shows the average share of Activities of Daily Living for which respondents reported having some difficulty, for those who were likely to be depressed versus those who were not. The right panel shows the rate of mortality in a two-year follow-up survey, separately for the group of individuals who were likely depressed versus those who were not at baseline. Except for the United States, the two waves used for comparison consist of the first (baseline) and second waves of each study. Displayed are 95 percent confidence intervals.

measure of functional impairment, we calculate the share of activities with which the respondents reported having at least some difficulty.

The depression indicator based on the screening questionnaires is strongly associated with impairments in activities of daily living. Panel A of Figure 2 compares the average of the share of activities for which the respondent had difficulty between those who show symptoms of depression and those who do not, controlling for age and gender.[5] In all countries, those who are likely depressed report having difficulty with at least twice as many activities, consistent with a literature documenting the

[5] We use linear regression to construct group averages for those who are predicted to be depressed versus those who are not (where an individual is predicted to be depressed if their score on the depression questionnaire surpasses the specified threshold). For the group that is not predicted to be depressed, we take the weighted average of the outcome (a measure of functional impairment) among all respondents who were not predicted to be depressed. For the group that is predicted to be depressed, we add to this average the coefficient on an indicator for predicted depression from a regression of the outcome on predicted depression, age bins, gender, and age bins interacted with gender. We use the

correlation between depression and functional disability in high-income contexts (Bruce 2001). For example, in China, the average respondent who shows symptoms of depression has difficulties with about 13 percent of the activities of daily living, compared to only 3 percent for the average respondent who does not exceed our depression score cutoff. Interestingly, the difference is largest in the United States. This correlation likely reflects a two-way causal relationship: restrictions in activities of daily living likely cause depression, while at the same time, depression may lower the ability and will to perform these activities.

Perhaps even more striking, depression symptoms are also associated with a higher likelihood of future death. We have panel data for older respondents at a two-year interval in several countries. Panel B of Figure 2 shows that elders who exhibited signs of depression in the earlier survey were more likely to have died two years later than nondepressed elders, after controlling for age and gender. For example, in Mexico, the two-year mortality rate for those not depressed at baseline was 3 percent, while the mortality rate for the depressed is significantly higher at 6 percent. The excess mortality among people with depression appears even more pronounced in the United States. These findings echo a literature on excess mortality among those with depressive symptoms, both among the elderly in rich countries (Adamson et al. 2005) and the elderly in low- and middle-income countries (Brandão et al. 2018). Once again, several factors could explain this relationship. Depression could be caused by poor health, which itself causes mortality. Alternatively, depression may accelerate death by leading to withdrawal from day-to-day activities that promote mobility and longevity, or by increasing the risk of suicide (World Health Organization 2018).

It is notable that, while old-age depression is more prevalent in low- and middle-income countries, the association of depression with poor outcomes (worse physical decline and increased risk of future death) appears in all countries in our sample, including the United States.

What Factors Are Associated with Depression in Old Age?

The literature has identified a number of factors associated with depression in old age (Blazer and Hybels 2014). Many of them could be both causes and effects.

Physical decline. As we document in Figure 2, the physical decline common to aging—falling, weight loss, frailty, or the inability to carry out daily activities such as bathing, walking, and household chores—is tightly linked to depression. This likely reflects a two-way causal relationship: poor physical health might be a cause of depression by reducing mobility and independence, or by causing physical pain or insomnia. Similarly, poor mental health could prevent one from maintaining physical fitness.

Lack of resources. Poverty and depression are often correlated (Ridley et al. 2020). Poverty can be an exacerbating factor for mental distress by exposing people to risk factors of depression, such as pollution, violence, low social status, poor sleeping

same methodology to control for age and gender when making other comparisons between subgroups in subsequent analyses.

conditions, and the inability to plan for economic shocks. Poverty also increases the risk of the physical health challenges described above by reducing healthcare expenditure. Finally, depression itself may be a risk factor for poverty through reduced labor supply and productivity, impaired decision-making, or discrimination by employers.

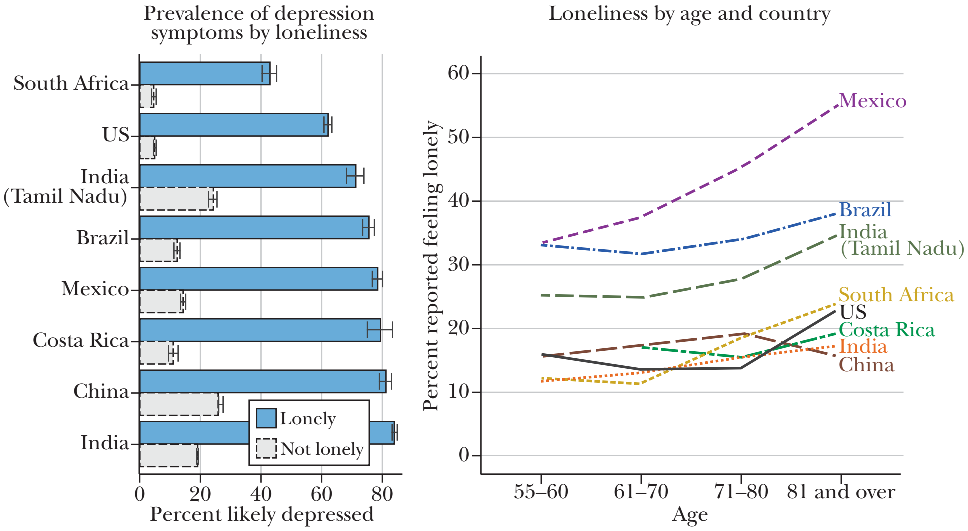
Lack of social support. Lack of social interactions and resulting feelings of loneliness are strong predictors of depression (Hawkley and Cacioppo 2010). One interpretation of this correlation is that loneliness and depression are two distinct but overlapping measures that capture different components of people's mental health. Alternatively, loneliness might cause depression, or vice versa. For instance, lack of social interactions can make people feel unsafe, generating a persistent fight-or-flight response which has myriad negative consequences: higher blood pressure, poor sleep quality, more negative social interactions, and a tendency to interpret social experiences negatively. Lack of companionship might lower one's sense of purpose and make it harder to enjoy life and deal with adverse shocks. For the elderly, lack of social support might be a particularly important cause of depression. As people age, their set of potential companions shrinks considerably due to untimely deaths of their loved ones, especially their spouses.

Panel A of Figure 3 shows a tight link between loneliness and depression. Elderly who report feeling lonely are much more likely to show symptoms of depression than those who do not: around 70 percent among the lonely versus 15 percent among the nonlonely.[6] One might assume that lack of social support is less of a problem in low- and middle-income countries because of differences in extended family living arrangements. While the fraction of elderly living alone in these countries still lags well behind Europe and the United States, it is rising as fertility drops and rural-to-urban migration increases (UN DESA Population Division 2017). However, feelings of loneliness appear to be at least as common in low- and middle-income countries as in the United States, as shown in panel B of Figure 3. The elderly report the highest rates of loneliness in Mexico, where 35 percent of people aged 61–70 reported feeling lonely a majority of the time—more than double the rate for the same age group in the United States. In most countries, loneliness is increasing with age and is around 10–30 percent for people aged 55–60, and 15–50 percent for those 80 or older.

The data described above provide evidence that the elderly living alone are particularly susceptible to loneliness and depression. However, these surveys are not particularly suitable to study the issues of the elderly living alone. While the demographic transition has increased the number of elderly living alone, the proportion

[6] Our main measure of loneliness is the answer to the question, "During the past week, did you feel lonely the majority of the time?" A downside of this measure is that it is subjective (as opposed to objective measures of social interactions) and thus may be vulnerable to social desirability bias due to stigma around the world "lonely." However, alternative measures of loneliness such as the UCLA Loneliness Scale may be less valid, particularly in low-income contexts (Mund et al. 2022). For example, questions such as "How often do you feel 'in tune' with the people around you?" may be interpreted very differently in different contexts.

Figure 3
Loneliness and Depression by Country



Source: Same as Figure 1.
Notes: The left panel shows the rate of depression among respondents separately by those who expressed feeling lonely most or all of the time in the past week and for those who did not, controlling for age and gender. For surveys in which self-reported loneliness is part of the depression index calculation (United States, China, India, Mexico, South Africa), we re-calculate depression scores excluding the response to the loneliness question and rescale accordingly. The right panel shows the percentage of respondents who expressed feeling lonely across each age group. For each country in our data except Malawi, we obtain a measure of loneliness from one direct survey question on loneliness. In some countries, the question asks respondents whether they felt lonely for the majority of the past week. We consider respondents who said "Yes" to be lonely. In other countries, the question asks how often they felt lonely. We consider respondents who expressed feeling lonely "Most or all of the time" to be lonely. For the Tamil Nadu data, the question was worded slightly differently and asked in reference to the previous week: "Do you often feel lonely?" Displayed are 95 percent confidence intervals.

remains low enough that their number is very small in any survey that does not focus on them. To get a glimpse of what the next decades may entail for elderly mental and physical health as the demographic transition increases the rate of elderly living entirely on their own, our Tamil Nadu survey focused explicitly on this very group of people.

The Tamil Nadu Aging Panel

To study the lives of the elderly with a particular focus on those living alone, we began conducting a large panel survey in the Indian state of Tamil Nadu in 2014. With a population of 76 million, Tamil Nadu lies in the southernmost part of the country. The Tamil Nadu Aging Panel is the result of a cooperation between the

Abdul Latif Jameel Poverty Action Lab (J-PAL) and the government of Tamil Nadu, cofunded by the United States National Institute of Health and the government of Tamil Nadu. The interviews are conducted by government surveyors, and the survey instruments are provided by the J-PAL team of researchers. The data is publicly available (Duflo et al. 2022).

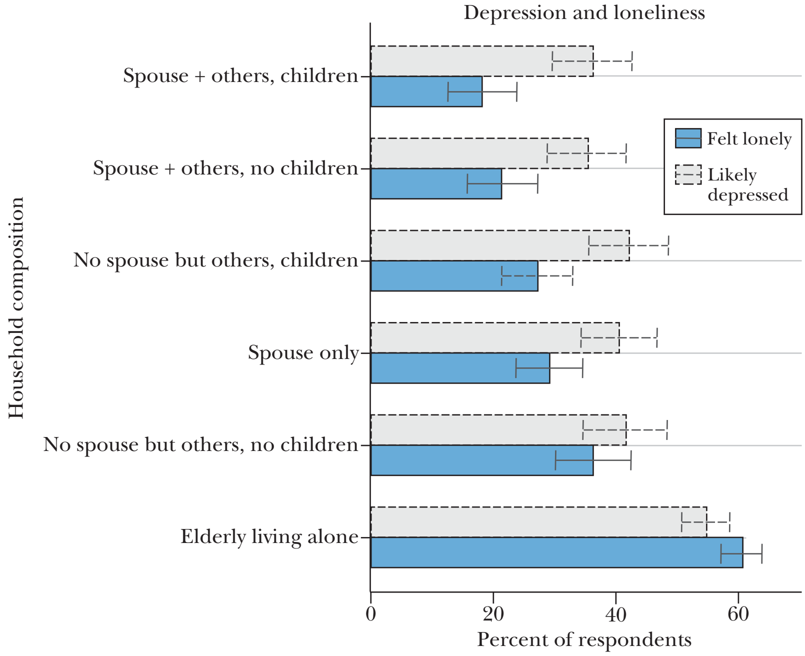
Compared to the previously available surveys, the Tamil Nadu Aging Panel has a wealth of additional information that allows us to look at a wider range of factors that can possibly explain elderly mental health. Additionally, it has a sufficient sample size for us to specifically focus on the issue of elderly living alone, which is poised to be a growing policy challenge around the world in the years ahead. Tamil Nadu, a fast-growing state, is a good bellwether for these trends: it has experienced a rapid demographic transition (the total fertility rate is currently only 1.8 children per woman), life expectancy growth (increasing from 66 years in 2002 to 74 years in 2019), and urbanization (48.5 percent of the population lives in an urban area, compared to a national average of 31 percent)—three factors likely to give rise to a rapid growth in the number of elderly living alone.

To obtain a sample frame for the Tamil Nadu Aging Panel, we first conducted a census of all households in each "enumeration area" (for example, a village hamlet). One key finding of the census was that 9 percent of those above 55 (and nearly 14 percent of women of that age) lived entirely alone. Those living alone tend to be older than those living with others. Furthermore, most people living alone are female and are often widows. The gender pattern may arise because women tend to be younger than their spouses and have a longer life expectancy, leading to a much higher fraction of females who are widows compared to the fraction among males. In our sample, for example, 54 percent of women above age 60 had a spouse who is now deceased compared to only 9 percent of men.

The high fraction of elderly living alone came as a surprise to our government partners: the expectation in India is that widows should live with relatives, usually their sons and daughters-in-law. Despite the demographic transition, most of the widows living alone (81 percent) in our sample had at least one child. But social norms seem to be changing. When we asked those with children who were still living alone, the most common reason for not living with their children was that they could not live with their son-in-law (50 percent), they did not want to be a burden on anyone (40 percent), and/or did not want to leave home (20 percent). Consistent with this finding, an earlier 2011 survey in Tamil Nadu found that approximately half of the elderly report a preference for living alone or with their spouse over living with their children or other relatives (Sathyanarayana et al. 2014). This shift in preferred living arrangement is proceeding rapidly in the developing world: the proportion of elderly living alone or with their spouse in India increased from 9 percent to 19 percent in just over a decade (Jadhav et al. 2013).

In light of this new policy concern, we decided to oversample the population of elderly living alone for our panel survey. The goal was to better understand the implications of living alone for physical and mental health, as well as to develop and test interventions that could be helpful to elderly living alone.

Figure 4
Social Interaction and Loneliness for the Elderly Living Alone in Tamil Nadu



Source: Data for Tamil Nadu come from the first wave of our own study that was conducted in the state of Tamil Nadu in 2019.
Notes: This figure shows the percentage of elderly in Tamil Nadu who report often feeling lonely as well as the percentage of the elderly who are likely to be depressed (as determined by whether their score on the depression index surpassed the specified threshold for at least mild depression), separately by household composition. We include controls for age and gender, following a similar methodology to the construction for Figures 2 and 3. First, we set the reference group to be Elderly Living with Others and calculate the weighted average of the outcome for that group. Then we obtain the coefficient of an indicator for Elderly Living Alone from a regression which includes controls for age and gender interacted. Displayed are 95 percent confidence intervals.

Elderly Living Alone, Isolation, and Mental Health

The elderly as a group in Tamil Nadu are quite likely to be depressed, as shown in Figure 1. Among elderly men, the share of likely depressed individuals increases from 30 percent for those 55–60 years of age to 43 percent for those over 80 years of age. Among elderly women, the shares are even higher: 37 percent for those 55–60 years of age and 49 percent for those over 80.

The elderly living alone are more likely to report feeling lonely, and the magnitude of this difference is striking. Figure 4 shows the percent of elders with various living arrangements who report often feeling lonely, controlling for age and gender. The elderly whose households include their spouse have the lowest loneliness rates: around 20–30 percent report feeling lonely. In contrast, this figure is 27–36 percent for the elderly who do not live with their spouses but do live with others, and 60 percent for the elderly living alone.

Figure 4 also shows higher rates of depression among the elderly living alone than for the elderly in other living arrangements. The elderly living alone are worse off on several other dimensions: controlling for age and gender, the elderly living alone have lower asset ownership, lower food security, and more cognitive impairment.

These findings are reminiscent of Chen and Drèze (1992), who 30 years ago examined the marginalization, poor health, and low well-being of widows in North India and made the point that this was an overlooked population. More recently, Srivastava et al. (2021) also find that living alone and widowhood are two highly significant predictors of poor mental health in the Longitudinal Aging Study in India dataset.

Poverty and Mental Health

We also use our panel data to document the link between lack of resources and depression among the elderly, consistent with evidence from other settings that poverty and depression are positively correlated (Ridley et al. 2020).

As seen in Figure 5, rates of depression fall steeply with expenditure. We estimate daily expenditure per person in a household using our survey data. Nearly 50 percent of elderly individuals in the lowest quintile of expenditure are likely to be depressed. By contrast, the rate is about 27 percent in the highest quintile. We also asked the elderly to assess their own financial status. Again, we find rates of depression fall steeply with financial well-being.

Of course, the underlying causal relationships are likely complex. However, the correlational evidence suggests that cash transfers and other financial support could help reduce depression rates among the elderly.

What Might Be Done about Elderly Mental Health in Low- and Middle-Income Countries?

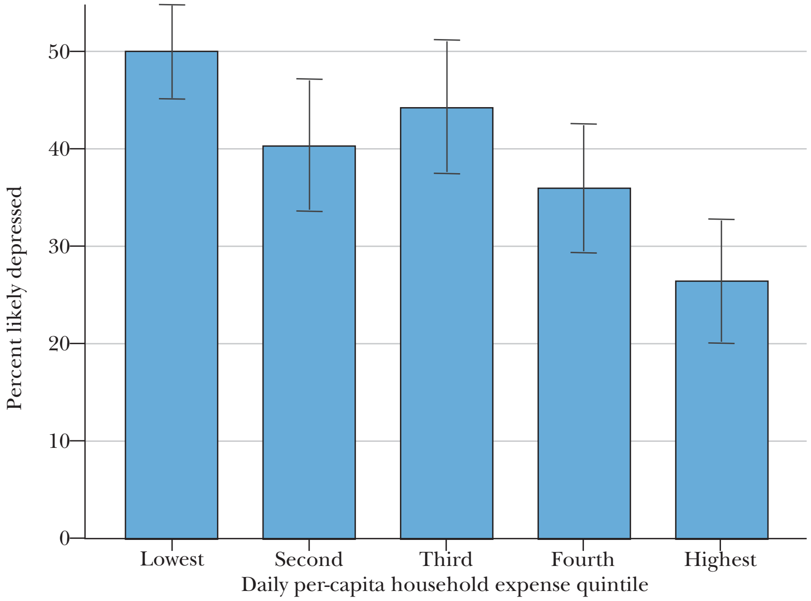
Research on the economic and psychological lives of the elderly poor has hardly scratched the surface so far. The high levels of psychological distress for this group warrants urgent attention from researchers, but given the largely descriptive and correlational evidence, it is too early to make strong policy prescriptions. Rather, our evidence at this stage, as well as past experiments in the general population, generates ideas for policies that might help improve the psychological well-being among the elderly. This section outlines some of these ideas in three areas: (1) poverty alleviation programs; (2) psychosocial interventions to reduce depression; and (3) interventions to improve physical health.

Cash Transfers and Old-Age Pensions

Cash transfers have been found to improve mental health in many settings with working-age adults (Ridley et al. 2020), and alleviating financial hardship could also improve mental health among elderly adults. Noncontributory old-age pensions

Figure 5

Depression Rates by Household Expenses and Financial Health



Source: Same as Figure 4.
Notes: This figure shows the prevalence of depression symptoms for respondents in Tamil Nadu, as a function of household expenses. We show the fraction likely depressed for each quintile of financial well-being, as measured by daily per-capita household expenses. An individual is tagged as likely depressed using the methodology described in Figure 1. The per-capita household expenses that correspond to each of the five quintiles are: [0, 29.7), [29.7, 44.3), [44.3, 60.2), [60.2, 93.4), [93.4, 328] Rs per day. All statistics are weighted by the inverse sampling probability. 95 percent confidence intervals are shown. Displayed are 95 percent confidence intervals.

(that is, regular cash transfers to the elderly), which are becoming increasingly popular across the world, offer a promising policy approach.

A growing body of evidence suggests that these pensions are effective at reducing symptoms of depression, including evaluations of such programs using quasi-experimental variation in China, Mexico, and Peru (Chen, Wang, and Busch 2019; Galiani, Gertler, and Bando 2016; Bando, Galiani, and Gertler 2020) and a randomized trial in Paraguay (Bando, Galiani, and Gertler 2022). A randomized trial of a year-long cash transfer program for older people in Nigeria found reductions in symptoms of depression after six months, though not after twelve (Alzua et al. 2020).

In preliminary results from on-going work in collaboration with the government of Tamil-Nadu (Banerjee et al. 2022), we find no significant effect on mental health of a government effort to deliver pensions to a random subset of elderly eligible for the Tamil Nadu Old Age pension but not receiving it. The Tamil Nadu Old Age pension is a noncontributory transfer of ₹1,000 per month (US$12 at market exchange rates, or $43 at purchasing power parity), reserved to individuals over age 60 who can demonstrate that they are "destitute" and cannot be supported by their family. Using data from our initial census, we identified households with elders who were not getting the pension but were likely eligible for it (data from our census suggest that over half of those likely eligible for the pension were not receiving it). We randomly divided the list in two and gave the lists to the government's Department of Economics and Statistics, which then transmitted the treatment group to the Department of Revenue for review. Not everybody was deemed eligible: by July 2022, 51 percent of the treatment list had received the pension while 18 percent of the comparable households in the control list had, which gives us a source of exogenous variation in pension receipt. We find no significant difference in mental health between the two groups, even when focusing on elderly living alone.

More work is needed on this topic: to our knowledge, the Paraguay study and ours are the only two randomized evaluations of a permanent government pension program, and they come to different conclusions. While the nonexperimental evidence to date argues that noncontributory cash transfer programs could help to improve mental health for the elderly, additional evidence on the mental health impact of pensions is needed.

Psychosocial Interventions to Reduce Depression

In recent decades, considerable progress has been made in effectively treating symptoms of depression by means of psychotherapy and pharmacotherapy. In addition, interventions to reduce social isolation and loneliness might help reduce depression among the elderly, though evidence in this area is just beginning to accumulate.

Therapy. A large body of evidence has shown that various forms of treatment—for example, cognitive behavioral therapy or the prescription of antidepression medication—effectively reduce symptoms of depression in many settings, including among the elderly. However, these treatments are typically unavailable in low-income contexts due to lack of resources: most notably, trained psychiatrists. To fill this void, simplified forms of therapy that can be administered by laypeople at a low cost have been developed and found often to be effective in the general population (Barbui et al. 2020). For instance, in Goa, India, a randomized trial found that delivering up to eight nonspecialist therapy sessions led to a reduction of 11 percentage points in symptoms of depression five years later (Bhat et al. 2022). Research on the elderly is much sparser. But also in Goa, a smaller-scale study of a similar intervention among the elderly was found to be effective in preventing depression (Dias et al. 2019).

When developing programs for scale, trade-offs in effectiveness may arise. In Tamil Nadu, light-touch, phone-delivered cognitive behavioral therapy that focused on problem solving and behavioral activation reduced functional impairment among women living alone three weeks after the conclusion of therapy (particularly the ability to carry out activities in social contexts). However, the effect had disappeared by three months, there were no effects on depressive symptoms at any time horizon, and men living alone did not benefit (and may have been made worse off) (McKelway et al. 2022). Compared to a one-time cash transfer of ₹1,000, this therapy program—even delivered by phone—was more expensive and less effective.

Therefore, a key question for the cost-effectiveness of therapy is whether it can be designed and implemented to yield sustained, long-term effects, and whether it can be embedded in existing government programs to reduce its costs. In nonelderly populations, some interventions have shown persistent effects for up to seven years (Baranov et al. 2020; Bhat et al. 2022). Among the elderly, the effects appear to fade faster, perhaps due to impaired memory, suggesting the importance of regular booster sessions as in Dias et al. (2019). One promising avenue of research is to train support persons (like family and neighbors) to provide simple therapy booster sessions, or by integrating therapy and boosters into regular health care or social welfare checks. For example, the Tamil Nadu government recently launched a program targeted at adults at risk for diseases like hypertension and diabetes: village health volunteers visit eligible adults' homes to conduct health tests and deliver medication. Such a program could expand to include training for the village health workers to provide some cognitive behavioral therapy as part of their visits. Another government organization, the Tamil Nadu Corporation for the Development of Women, hires community resource persons from women's self-help groups to initiate programs in communities on a range of issues, including food, sanitation, and health. These women could be trained to provide community members with therapy, both initial sessions and boosters, as part of their work.

Beyond treating the currently depressed, interventions that help to prevent future episodes of depression could be valuable. Barker et al. (2022) find that group therapy in Ghana reduced future symptoms of depression even for people who were not depressed at baseline, but who were likely to become depressed based on baseline information. The idea is that therapy teaches people how to deal with future shocks or other triggers by understanding which activities might help improve their mood. Evidence from high-income countries shows promise in preventative interventions and suggests targeting at-risk individuals (like elderly living alone, or widows). This evidence also suggests that encouraging the elderly to engage in social and physical activities might be effective in preventing depression (Park, Han, and Kang 2014).

Fostering social interactions. If social isolation and loneliness are key drivers of poor mental health for the elderly, it may be difficult for intermittent therapy to overcome such ongoing conditions. Even nonlonely individuals tend to underestimate the benefits of social interactions (Epley and Schroeder 2014); loneliness can be self-reinforcing by changing the way people think about and value social interactions and by changing people's mood, perceptions, and behavior toward others (Cacioppo and Patrick 2009). As a result, demand for social interactions might be inefficiently low, thus providing scope for interventions that foster social interactions.

At present, there is little evidence on well-powered interventions to tackle social isolation and loneliness among the elderly in any context, and especially in low-income countries (Masi et al. 2011). Some possibilities should be explored. For example, many of the elderly have family members or friends with whom they could connect more frequently (even if they do not live together). However, family providers—in particular, young women, on whom the burden of caring for the elderly often falls—also need to be supported to avoid harmful effects on their labor supply or mental health.

Opportunities for connections outside of the family should also be explored. In India, women typically leave their natal villages upon marriage and face restrictions on their physical mobility once they are married, leading to substantial risk of loneliness. Indeed, research on the social networks of young married women in India reveals high amounts of isolation (Andrew et al. 2020). Socializing outside of the immediate family can improve mental health. For example, senior citizen clubs and activities may encourage socialization and reduce loneliness but are mostly absent in villages in low- and middle-income countries. Some evidence from high-income countries suggests that these types of opportunities can improve physical and mental health, including randomized control trials involving activities such as dancing, walking, and tai chi (Rogers et al. 2009).

Even relatively light-touch interventions can improve psychological well-being, such as providing phones (and teaching people how to use them) or phone credit to enable increased communication (for an experiment in Ghana, see Annan and Archibong 2022). Similarly, employing laypeople to call the elderly regularly during the pandemic reduced depression among the elderly in Texas (Kahlon et al. 2021). To conduct our randomized evaluation of phone-based therapy in Tamil Nadu, where all activities had to happen in a socially distanced way due to COVID-19, we delivered cell phones and trained older people in their use. Even though many recipients had never had a phone before, the participants used their phones extensively and continued to do so after the experiment ended. Making phones available to the elderly might be a promising intervention to test at scale.

It may also be productive to combine interventions that both increase the demand for social interactions, such as cognitive behavioral therapy, and those that increase the supply; simply increasing the supply of social interactions might not be enough to improve outcomes, because without other support, a lonely person might not be in the right mindset to take advantage of these opportunities.

Restoring dignity and sense of purpose. Finally, interventions to restore dignity and a sense of purpose could be important. An elderly person who was once a respected member of the community—whether as someone who was in charge of raising children, the lead decision-maker in the household, or the family breadwinner—may feel a loss of purpose or dignity as they lose those responsibilities with aging. Providing opportunities for elders to maintain or rebuild their (perceived) ability to contribute to their communities could strengthen purpose, dignity, and thus mental health. Offering work to nonelderly refugees has meaningful benefits beyond the cash value of this work, including reduced depression (Hussam et al. 2022). Similar benefits could be achieved for the elderly in low-income contexts through interventions that involve them in childcare or work opportunities. Even if the productivity of the elderly is low, significant mental health benefits can arise from remaining active and engaged and from feeling a sense of purpose.

Improving Mental Health by Improving Physical Health

Physical and mental health are positively correlated, and it seems plausible that the causality between them runs in both directions. Here, we focus on possible steps for improving physical health, as a possible driver of mental health.[7]

Reducing physical pain. Significant physical pain is nearly universal for the elderly, and the experience of physical pain is strongly associated with poor mental health (Bair et al. 2003). Pain can often be addressed through the treatment of underlying health conditions (like arthritis or dental decay) or through psychological interventions. In high-income countries, cognitive behavioral therapy for chronic pain has shown some promise (Ehde, Dillworth, and Turner 2014).

Improving functional abilities. Loss of hearing, vision, or mobility impedes people's ability to communicate and socialize, as well as to carry out activities of daily living. Correlational evidence suggests that this may increase their sense of isolation and deteriorate mental health (Marmamula et al. 2021). However, in low-income settings, where access to healthcare is limited, the elderly are often unaware that they have functional limitations—and that these limitations can be treated. In our Tamil Nadu data, nearly 50 percent of elders were evaluated as hearing impaired, but under 30 percent of elders actually report experiencing hearing loss. Similarly, about 45 percent are diagnosed with visual impairment due to cataracts, but under 30 percent report having this condition. It should be a priority to provide widespread access to affordable, high-quality devices to mitigate specific functional impairment—such as hearing aids, eyeglasses, and walkers—for the elderly in low-income settings. Providing eyeglasses has been shown to increase work productivity for nonelderly populations (Reddy et al. 2018), and providing the elderly in China with hearing aids has been found to improve life satisfaction (Ye et al. 2022). For the elderly, such benefits could involve being able to leave their homes again on their own, visit their friends, or just to enjoy sight itself, all of which may improve their mental health.

Better management of chronic conditions. While chronic health conditions are on the rise, many remain undetected: in our Tamil Nadu data, for example, over 40 percent of the elderly were diagnosed with diabetes but less than half of those knew they had the condition; similarly, over 60 percent had hypertension but less than one-third of those knew about it. Regular check-ups may be particularly valuable—whether carried out through "health camps" organized close to people's homes or through

[7] Even though cognitive impairment (for example, in the form of dementia) is strongly associated with depression, we do not focus on it here because of lack of evidence on effective interventions to combat cognitive decline. However, dementia of the elderly is also correlated with depression in their caregiver in the family, suggesting potentially benefits from identifying and supporting families with an elderly person who experiences dementia (Dias et al. 2008).

at-home visits for the immobile elderly, who may also be among the most impaired and vulnerable. Tamil Nadu has recently launched a "health care at your doorstep" scheme, where frontline health workers visit households to diagnose chronic health conditions and then follow up with those diagnosed to get them medication. While such schemes seem promising, there is little work evaluating their impact on physical and mental health, and none in developing countries.[8] Perhaps mobile phones and cheap diagnostic tools along with machine learning techniques could be used both for current diagnoses and to influence future testing. Adhering to medication schedules can also be a challenge for the elderly. Thus, developing and using technologies to aid, encourage, and remind elderly adults in managing their medications, such as daily reminders and specially designed pill bottles that help keep count of medications, could complement traditional approaches like regular home visits to supply drugs.

Increase physical activity. Randomized control trials find exercise to be moderately effective at reducing depression among the elderly (Bigarella et al. 2022). In nonelderly populations, small incentives to walk have been shown to be effective at increasing exercise and mental health (Aggarwal, Dizon-Ross, and Zucker 2022), but it is not clear that they would be appropriate in a population with low mobility. Group activities, perhaps in the form of exercise classes, could both increase socialization and lead to some exercise. These types of activities and their benefits for the elderly in poor countries are highly promising and merit more careful research.

Improve sleep. Physical exercise could also help with poor sleep, another potential driver of cognitive decline and depression. The ability to sleep soundly declines substantially with age, which has been linked to cognitive decline (Mander, Winer, and Walker 2017). Helping older poor people sleep better could improve their wellbeing and mental health. But very little research exists on how to do this, especially in low- and middle-income countries. Cognitive behavioral therapy for insomnia has been shown to be effective in improving sleep quality in high-income nonelderly samples (Trauer et al. 2015).

Many of these steps could potentially build on each other: for example, efforts to improve physical health can help sleep, as well as management of chronic diseases like diabetes. Better health diagnoses might benefit people in many ways not mentioned here. Importantly, estimates of the gains from improving physical health should be expanded to include the corresponding gains to mental health as well.

Conclusion

More than 500 million people over 60 years of age live in low- and middle-income countries (UN DESA Population Division 2019), but the issues of their

[8] Liimatta et al. (2019) find positive impacts of home visits for adults above 75 on physical health and depression in Finland.

mental health and well-being are largely left unaddressed and unstudied. In the Sustainable Development Goals promulgated by the United Nations, across the 17 goals and 169 targets, the elderly are specifically mentioned only three times, and always as part of a list including many others—as in "those in vulnerable situations, women, children, persons with disabilities and older persons." As the share of older people in low- and middle-income countries continues to rise, efforts to improve social welfare will require paying more explicit attention to this group.

The elderly in low- and middle-income countries appear to be particularly vulnerable to poor mental health. The decline in capacities that comes with aging need not entail a directly corresponding decline in well-being and mental health: indeed, in the United States, we find that prevalence of symptoms of depression does not increase with age and is relatively low overall. In contrast, we find stark increases in depressive symptoms at older ages in several low- and middle-income countries. Interventions to improve mental health may well turn out to be very cost-effective ways to improve quality years of life, both because mental health is a key component of well-being and because improving mental health might have additional benefits for physical health and even survival. Although we have directed some attention in this paper to the elderly living alone, it is still the case that most elderly in low- and middle-income countries do not live alone; therefore, interventions to improve the mental health of the elderly may also have positive spillovers on the well-being of those charged with caring for them.

In this essay, we discussed potential avenues to improve the mental health of the elderly. Some seek to lower depression directly (for example, via therapy), while others target the physical, economic, or social root causes of depression. Much work remains to be done: most of these ideas have not been tested, and almost none have been implemented at scale. Finally, we acknowledge that this is just a small part of a much larger agenda on mental health in low- and middle-income countries, which should encompass other demographic groups, such as women or adolescents, and other mental health conditions, such as anxiety or post-traumatic stress disorder.

■ *Our deep thanks go to the J-PAL team in Tamil Nadu for many years of wonderful work that is reflected in this paper: Ankit Agarwal, Gunjita Gupta, Nikhil Kanakamedala, Murali Mallikarjunan, Sohaib Nasim, Cyrus Graham Reginald, and A. R. Selva Swetha. We also thank Julia Crocco, Catriona Farquharson, Maxwell Grozovsky, Mikey Jarrell, Laura Lahoz González, Laura Stillwell, and Jenny Wang for excellent research assistance. We thank Miriam Sequeira for her insights. Financial assistance is provided by the NIA under grant number P01AG005842, and by the Tamil Nadu Government under the J-PAL-TNAGE partnership. Many thanks to Aparna Krishna, of J-PAL, and Mr. Krishnan, of the Tamil Nadu government, for getting this partnership off the ground and overseeing the start of this project.*

References

Adamson, Joy A., Gill M. Price, Elizabeth Breeze, Christopher J. Bulpitt, and Astrid E. Fletcher. 2005. "Are Older People Dying of Depression? Findings from the Medical Research Council Trial of the Assessment and Management of Older People in the Community." *Journal of the American Geriatrics Society* 53 (7): 1128–32.

Aggarwal, Shilpa, Rebecca Dizon-Ross, and Ariel Zucker. 2022. "Designing Incentives for Impatient People: An RCT Promoting Exercise to Manage Diabetes." Unpublished.

Alzua, Maria Laura, Natalia Cantet, Ana C. Dammert, and Damilola Olajide. 2020. "The Wellbeing Effects of an Old Age Pension: Experimental Evidence for Ekiti State in Nigeria." Paper presented at 2020 Agricultural and Applied Economics Association Annual Meeting, Kansas City, MO, July 26–28.

American Psychiatric Association. 2022. "Depressive Disorders." *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*. 5th ed. Washington, DC: American Psychiatric Association Publishing.

Andrew, Alison, Orazio Attanasio, Britta Augsburg, Jere Behrman, Monimalika Day, Pamela Jervis, Costas Meghir, and Angus Phimister. 2020. "Mothers' Social Networks and Socioeconomic Gradients of Isolation." NBER Working Paper 28049.

Annan, Francis, and Belinda Archibong. 2022. "The Value of Communication for Mental Health." Brookings Global Working Paper 177.

Bair, Matthew J., Rebecca L. Robinson, Wayne Katon, and Kurt Kroenke. 2003. "Depression and Pain Comorbidity: A Literature Review." *Archives of Internal Medicine* 163 (20): 2433–45.

Banerjee, Abhijit, Esther Duflo, Erin Grela, Madeline McKelway, Frank Schilbach, Garima Sharma, and Girija Vaidyanathan. 2023. "Replication data for: Depression and Loneliness among the Elderly in Low- and Middle-Income Countries." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E185121V1.

Banerjee, Abhijit, Esther Duflo, Erin Grela, Madeline McKelway, Frank Schilbach, Garima Sharma, and Girija Vaidyanathan. 2022. "The Causal Effects of Old Age Pensions." AEA RCT Registry. https://www.socialscienceregistry.org/trials/4140.

Bando, Rosangela, Sebastian Galiani, and Paul Gertler. 2020. "The Effects of Noncontributory Pensions on Material and Subjective Well-Being." *Economic Development and Cultural Change* 68 (4): 1233–55.

Bando, Rosangela, Sebastian Galiani, and Paul Gertler. 2022. "Another Brick on the Wall: On the Effects of Non-contributory Pensions on Material and Subjective Well Being." *Journal of Economic Behavior and Organization* 195: 16–26.

Baranov, Victoria, Sonia Bhalotra, Pietro Biroli, and Joanna Maselko. 2020. "Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial." *American Economic Review* 110 (3): 824–59.

Barbui, Corrado, Marianna Purgato, Jibril Abdulmalik, Ceren Acarturk, Julian Eaton, Chiara Gastaldon, Oye Gureje, et al. 2020. "Efficacy of Psychosocial Interventions for Mental Health Outcomes in Low-Income and Middle-Income Countries: An Umbrella Review." *The Lancet Psychiatry* 7 (2): 162–72.

Barker, Nathan, Gharad Bryan, Dean Karlan, Angela Ofori-Atta, and Christopher Udry. 2022. "Cognitive Behavioral Therapy among Ghana's Rural Poor Is Effective Regardless of Baseline Mental Distress." *American Economic Review: Insights* 4 (4): 527–45.

Berkman, Lisa. 2023. "Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community is South Africa [HAALSI]: Agincourt, South Africa, 2015–2022." Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR36633.

Bhat, Bhargav, Jonathan de Quidt, Johannes Haushofer, Vikram H. Patel, Gautam Rao, Frank Schilbach, and Pierre-Luc P. Vautrey. 2022. "The Long-Run Effects of Psychotherapy on Depression, Beliefs, and Economic Outcomes." NBER Working Paper 30011.

Bigarella, Lucas Goldmann, Vinicius Remus Ballotin, Lucas Ferrazza Mazurkiewicz, Ana Carolina Ballardin, Dener Lizot Rech, Roberto Luis Bigarella, and Luciano da Silva Selistre. 2022. "Exercise for Depression and Depressive Symptoms in Older Adults: An Umbrella Review of Systematic Reviews and Meta-analyses." *Aging and Mental Health* 26 (8): 1503–13.

Blanchflower, David G. 2021. "Is Happiness U-shaped Everywhere? Age and Subjective Well-Being in 145 Countries." *Journal of Population Economics* 34: 575–624.

Blazer, Dan G. and Celia F. Hybels. 2014. "Depression in Later Life: Epidemiology, Assessment, Impact, and Treatment." In *Handbook of Depression*, edited by Ian H. Gotlib and Constance L. Hammen, 429–47. New York: Guilford Press.

Bloom, David E., T. V. Sekher, and Jinkook Lee. 2021. "Longitudinal Aging Study in India (LASI): New Data Resources for Addressing Aging in India." *Nature Aging* 1: 1070–72.

Brandão, Glauber Sá, Luís Vicente Franco Oliveira, Glaudson Sá Brandão, Anderson Soares Silva, Antônia Adonis Callou Sampaio, Jessica Julioti Urbano, Alyne Soares, et al. 2018. "Effect of a Home-Based Exercise Program on Functional Mobility and Quality of Life in Elderly People: Protocol of a Single-Blind, Randomized Controlled Trial." *Trials* 19: 684.

Bruce, Martha L. 2001. "Depression and Disability in Late Life: Directions for Future Research." *American Journal of Geriatric Psychiatry* 9 (2): 102–12.

Bugliari, Delia, Joanna Carroll, Orla Hayden, Jessica Hayes, Michael D. Hurd, Adam Karabatakis, Regan Main, Colleen M. McCullough, Erik Meijer, Michael B. Moldoff, Philip Pantoja, Susann Rohwedder, Patricia St. Clair. 2022. Health and Retirement Study, (RAND HRS Longitudinal File 2018 (V2)) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI.

Bugliari, Delia, Joanna Carroll, Orla Hayden, Jessica Hayes, Michael Hurd, Adam Karabatakis, Regan Main et al. 1992–2018. "RAND HRS Longitudinal File 2018 V1." RAND Center for the Study on Aging. https://www.rand.org/content/dam/rand/www/external/labor/aging/dataprod/randhrs1992_2018v1.pdf (accessed July 1, 2022).

Cacioppo, John T. and William Patrick. 2009. *Loneliness: Human Nature and the Need for Social Connection.* New York: W. W. Norton & Company.

Chen, Marty, and Jean Drèze. 1992. "Widows and Health in Rural North India." *Economic and Political Weekly* 27 (43/44): WS81–92.

Chen, Xi, Tianyu Wang, and Susan H. Busch. 2019. "Does Money Relieve Depression? Evidence from Social Pension Expansions in China." *Social Science and Medicine* 220: 411–20.

Dias, Amit, Fredric Azariah, Stewart J. Anderson, Miriam Sequeira, Alex Cohen, Jennifer Q. Morse, Pim Cuijpers, Vikram Patel, and Charles F. Reynolds III. 2019. "Effect of a Lay Counselor Intervention on Prevention of Major Depression in Older Adults Living in Low- and Middle-Income Countries: A Randomized Clinical Trial." *JAMA Psychiatry* 76 (1): 13–20.

Dias, Amit, Michael E. Dewey, Jean D'Souza, Rajesh Dhume, Dilip D. Motghare, K. S. Shaji, Rajiv Menon, Martin Prince, and Vikram Patel. 2008. "The Effectiveness of a Home Care Program for Supporting Caregivers of Persons with Dementia in Developing Countries: A Randomised Controlled Trial from Goa, India." *PloS ONE* 3 (6): e2333.

Duflo, Esther, Abhijit Banerjee, Madeline McKelway, Frank Schilbach, Garima Sharma, and Girija Vaidyanathan. 2022. "Tamil Nadu Aging Panel." https://doi.org/10.7910/DVN/SXEYFW, Harvard Dataverse, V7, UNF:6:L98lZYlNOvESyoxsLYoJJw==[fileUNF]

Eaton, W. W., M. Kramer, J. C. Anthony, A. Dryman, S. Shapiro, and B. Z. Locke. 1989. "The Incidence of Specific DIS/DSM-III Mental Disorders: Data from the NIMH Epidemiologic Catchment Area Program." *Acta Psychiatrica Scandinavica* 79 (2): 163–78.

Ehde, Dawn M., Tiara M. Dillworth, and Judith A. Turner. 2014. "Cognitive-Behavioral Therapy for Individuals with Chronic Pain: Efficacy, Innovations, and Directions for Research." *American Psychologist* 69 (2): 153–66.

Epley, Nicholas, and Juliana Schroeder. 2014. "Mistakenly Seeking Solitude." *Journal of Experimental Psychology: General* 143 (5): 1980–99.

Galiani, Sebastian, Paul Gertler, and Rosangela Bando. 2016. "Non-contributory Pensions." *Labour Economics* 38: 47–58.

Giuntella, Osea, Sally McManus, Redzo Mujcic, Andrew J. Oswald, Nattavudh Powdthavee, and Ahmed Tohamy. 2023. "The Midlife Crisis." *Economica* 90 (357): 65–110.

Hawkley, Louise C., and John T. Cacioppo. 2010. "Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms." *Annals of Behavioral Medicine* 40 (2): 218–27.

Hussam, Reshmaan, Erin M. Kelley, Gregory Lane, and Fatima Zahra. 2022. "The Psychosocial Value of Employment: Evidence from a Refugee Camp." *American Economic Review* 112 (11): 3694–3724.

Institute for Health Metrics and Evaluation (IHME). 2019a. "Global Burden of Disease Study 2019 (GBD 2019) Data Input Sources Tool." University of Washington. https://ghdx.healthdata.org/gbd-2019/data-input-sources (accessed June 1, 2022).

Institute for Health Metrics and Evaluation (IHME). 2019b. "Global Burden of Disease Study 2019 (GBD 2019) Results Tool." University of Washington. https://vizhub.healthdata.org/gbd-results/ (accessed August 10, 2022).

Jadhav, Apoorva, K. M. Sathyanarayana, Sanjay Kumar, and K. S. James. 2013. "Living Arrangements of the Elderly in India: Who Lives Alone and What Are the Patterns of Familial Support?" https://www.pop.upenn.edu/sites/www.pop.upenn.edu/files/PAA_Jadhav%202013_apoorva_PDF.pdf.

Kahlon, Maninder K., Nazan Aksan, Rhonda Aubrey, Nicole Clark, Maria Cowley-Morillo, Elizabeth A. Jacobs, Rhonda Mundhenk, Katherine R. Sebastian, and Steven Tomlinson. 2021. "Effect of Layperson-Delivered, Empathy-Focused Program of Telephone Calls on Loneliness, Depression, and Anxiety among Adults during the COVID-19 Pandemic: A Randomized Clinical Trial." *JAMA Psychiatry* 78 (6): 616–22.

Kahneman, Daniel, and Alan B. Krueger. 2006. "Developments in the Measurement of Subjective Well-Being." *Journal of Economic Perspectives* 20 (1): 3–24.

Kohler, Hans-Peter, Susan C. Watkins, Jere R. Behrman, Philip Anglewicz, Iliana V. Kohler, Peter Fleming, Rebecca L. Thornton et al. 2013. "Cohort Profile: The Malawi Longitudinal Study of Families and Health (MLSFH)." Population Studies Center, University of Pennsylvania, Working Paper 2013-06. http://repository.upenn.edu/psc_working_papers/46.

Liimatta, Heini, Pekka Lampela, Pirjo Laitinen-Parkkonen, and Kaisu H. Pitkala. 2019. "Effects of Preventive Home Visits on Health-Related Quality-of-Life and Mortality in Home-Dwelling Older Adults." *Scandinavian Journal of Primary Health Care* 37 (1): 90–97.

Lima-Costa, M. Fernanda, Fabíola Bof de Andrade, Paulo Roberto Borges de Souza Jr., Anita Liberalesso Neri, Yeda Aparecida de Oliveira Duarte, Erico Castro-Costa, Cesar de Oliveira. 2018. "The Brazilian Longitudinal Study of Aging (ELSI-Brazil: Objectives and Design." *American Journal of Epidemiology* 187 (7): 1345–53. doi: 10.1093/aje/kwx387.

Mander, Bryce A., Joseph R. Winer, and Matthew P. Walker. 2017. "Sleep and Human Aging." *Neuron* 94 (1): 19–36.

Marmamula, Srinivas, Thirupathi Reddy Kumbham, Satya Brahmanandam Modepalli, Navya Rekha Barrenkala, Ratnakar Yellapragada, and Rahul Shidhaye. 2021. "Depression, Combined Visual and Hearing Impairment (Dual Sensory Impairment): A Hidden Multi-morbidity among the Elderly in Residential Care in India." *Nature: Scientific Reports* 11: 16189.

Masi, Christopher M., Hsi-Yuan Chen, Louise C. Hawkley, and John T. Cacioppo. 2011. "A Meta-analysis of Interventions to Reduce Loneliness." *Personality and Social Psychology Review* 15 (3): 219–66.

Maxim, L. Daniel, Ron Niebo, and Mark J. Utell. 2014. "Screening Tests: A Review with Examples." *Inhalation Toxicology* 26 (13): 811–28.

McKelway, Madeline, Abhijit Banerjee, Erin Grela, Frank Schilbach, Garima Sharma, Miriam Sequeira, Girija Vaidyanathan, and Esther Duflo. 2022. "Impacts of Cognitive Behavioral Therapy and Cash Transfers on Depression and Impairment of Elderly Living Alone: A Randomized Trial in India." Unpublished.

Mund, Marcus, Marlies Maes, Pia M. Drewke, Antonia Gutzeit, Isabel Jaki, and Pamela Qualter. 2022. "Would the Real Loneliness Please Stand Up? The Validity of Loneliness Scores and the Reliability of Single-Item Scores." *Assessment*. https://doi.org/10.1177/10731911221077227.

Park, Seong-Hi, Kuem Sun Han, and Chang-Bum Kang. 2014. "Effects of Exercise Programs on Depressive Symptoms, Quality of Life, and Self-Esteem in Older People: A Systematic Review of Randomized Controlled Trials." *Applied Nursing Research* 27 (4): 219–26.

Reddy, Priya Adhisesha, Nathan Congdon, Graeme MacKenzie, Parikshit Gogate, Qing Wen, Catherine Jan, Mike Clarke et al. 2018. "Effect of Providing Near Glasses on Productivity among Rural Indian Tea Workers with Presbyopia (PROSPER): A Randomised Trial." *The Lancet Global Health* 6 (9): e1019–27. https://doi.org/10.1016/S2214-109X(18)30329-2.

Ridley, Matthew, Gautam Rao, Frank Schilbach, and Vikram Patel. 2020. "Poverty, Depression, and Anxiety: Causal Evidence and Mechanisms." *Science* 370 (6522): eaay0214.

Rogers, Carol E., Linda K. Larkey, and Colleen Keller. 2009. "A Review of Clinical Trials of Tai Chi and Qigong in Older Adults." *Western Journal of Nursing Research* 31 (2): 245–79.

Rosero-Bixby, Luis, Xenia Fernández, and William H. Dow. 2013. "CRELES: Costa Rican Longevity and Healthy Aging Study, 2005 (Costa Rica Estudio de Longevidad y Envejecimiento Saludable)." Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR26681.v2.

Salk, Rachel H., Janet S. Hyde, and Lyn Y. Abramson. 2017. "Gender Differences in Depression in Representative National Samples: Meta-analyses of Diagnoses and Symptoms." *Psychological Bulletin* 143 (8): 783–822.

Sathyanarayana, K. M., Lekha Subaiya, S. Ravichandran, and Supriya Verma. 2014. "The Status of the

Elderly in Tamil Nadu, 2011." United Nations Population Fund. http://www.isec.ac.in/Tamil%20 Nadu.pdf.

Sonnega, Amanda, Jessica D. Faul, Mary Beth Ofstedal, Kenneth M. Langa, John W.R. Phillips, and David R. Weir. 2014. "Cohort Profile: The Health and Retirement Study (HRS)." *International Journal of Epidemiology* 43 (2): 576–85.

Srivastava, Shobhit, Paramita Debnath, Neha Shri, and T. Muhammad. 2021. "The Association of Widowhood and Living Alone with Depression among Older Adults in India." *Nature: Scientific Reports* 11 (21641).

Thornicroft, Graham, Somnath Chatterji, Sara Evans-Lacko, Michael Gruber, Nancy Sampson, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi et al. 2017. "Undertreatment of People with Major Depressive Disorder in 21 Countries." *British Journal of Psychiatry* 210 (2): 119–24.

Trauer, James M., Mary Y. Qian, Joseph S. Doyle, Shantha M. W. Rajaratnam, and David Cunnington. 2015. "Cognitive Behavioral Therapy for Chronic Insomnia: A Systematic Review and Meta-analysis." *Annals of Internal Medicine* 163 (3): 191–204. https://doi.org/10.7326/M14-2841.

UN DESA (United Nations Department of Economic and Social Affairs), Population Division. 2017. *Living Arrangements of Older Persons: A Report on an Expanded International Dataset.* New York: United Nations.

UN DESA (United Nations Department of Economic and Social Affairs), Population Division. 2019. *World Population Ageing 2019: Highlights.* United Nations: New York.

Vilagut, Gemma, Carlos G. Forero, Gabriela Barbaglia, and Jordi Alonso. 2016. "Screening for Depression in the General Population with the Center for Epidemiologic Studies Depression (CES-D): A Systematic Review with Meta-analysis." *PloS ONE* 11 (5): e0155431.

Wong, Rebeca, Alejandra Michaels-Obregon, and Alberto Palloni. 2017. "Cohort Profile: The Mexican Health and Aging Study (MHAS)." *International Journal of Epidemiology* 46 (2): e2(1–10). https://doi.org/10.1093/ije/dyu263.

World Health Organization. 2018. *Guidelines for the Management of Physical Health Conditions in Adults with Severe Mental Disorders.* Geneva: World Health Organization.

World Health Organization. 2021. *Mental Health ATLAS 2020.* Geneva: World Health Organization.

Yang, Yang. 2007. "Is Old Age Depressing? Growth Trajectories and Cohort Variations in Late-Life Depression." *Journal of Health and Social Behavior* 48 (1): 16–32.

Ye, Xin, Dawei Zhu, Siyuan Chen, Xuefeng Shi, Rui Gong, Juncheng Wang, Huibin Zuo, and Ping He. 2022. "Effects of Providing Free Hearing Aids on Multiple Health Outcomes among Middle-Aged and Older Adults with Hearing Loss in Rural China: A Randomized Controlled Trial." *BMC Medicine* 20: 124.

Zhao, Yaohui, Yisong Hu, James P. Smith, John Strauss, and Gonghuan Yang. 2014. Cohort Profile: The China Health and Retirement Longitudinal Study (CHARLS)." *International Journal of Epidemiology* 43 (1): 61–68. https://doi.org/10.1093/ije/dys203.

Journal of Economic Perspectives—Volume 37, Number 2—Spring 2023—Pages 203–230

An Introductory Guide to Event Study Models

Douglas L. Miller

The event study model is a powerful econometric tool used for the purpose of estimating dynamic treatment effects. One of its most appealing features is that it creates a built-in graphical summary of results. In one of the earliest papers in labor economics to use an event study model, Jacobson, LaLonde, and Sullivan (1993) sought to estimate the loss of income after being displaced from a job. Figure 1 reproduces a graph from that paper.

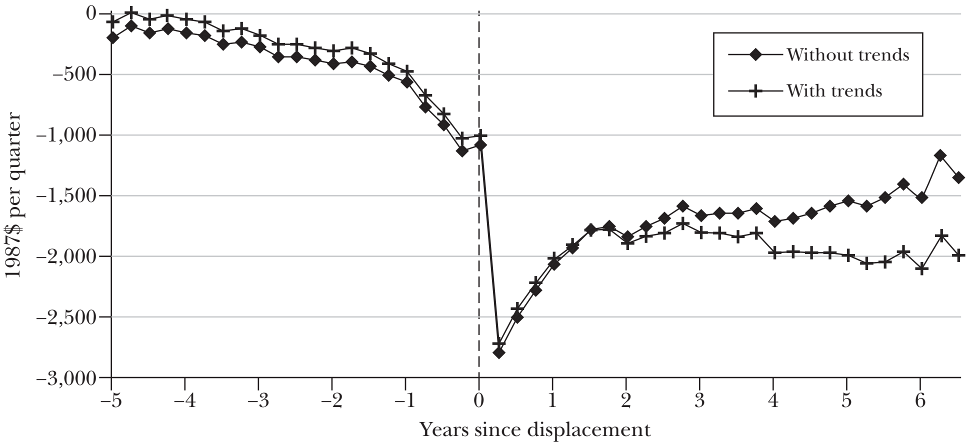
The *x*-axis is measured in "event time," meaning that for each person, the time of job displacement is treated as zero. The time-zero event is often referred to as the "treatment"—that is, the event or policy that changed what otherwise would have happened. The *y*-axis of the picture shows income for each period relative to a baseline comparison period. In this example, the baseline is more than five years prior to the job displacement.

The change after the event time of zero is the key takeaway from an event study picture, but the picture also reveals other rich patterns of behavior. For example, it also shows patterns before the event. Ideally, we hope that the line before the event is trendless, and deviations from that pattern alert us to a potential problem with our model; in particular, a trend suggests that the treatment may have been expected or that other factors are in play. In Figure 1, we see a modest deterioration in earnings in advance of the layoffs. This may reflect the presence of third factors—say, perhaps declines in demand for the output of a certain industry—that affect earnings prior to the event and that ultimately contribute to the displacement.

■ *Douglas L. Miller is Professor of Economics and Public Policy, Cornell University, Ithaca, New York. His email address is dlm336@cornell.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.203.

Figure 1
An Event Study Example: Loss of Income after Being Displaced from a Job



Source: Jacobson, LaLonde, and Sullivan (1993).
Note: Figure reproduced from Jacobson, LaLonde, and Sullivan (1993). The x-axis is measured in "event time." The y-axis show income for each period relative to a baseline comparison period more than five years prior to the job discplacement.

Alternatively, if displacement was anticipated and resulted in discouragement, this could lead to a pre-event trend through reductions in labor supply. The figure shows a modest pre-event trend, but also shows a sharp drop in earnings at the time of displacement, followed by a bounce back over the next two years that levels off at an earnings decline of about $500 to $1,000 per quarter compared to the pre-event level.

Event study models in economics started with finance applications: for a survey of earlier event studies in finance, see MacKinlay (1997). His earliest example is Dolley (1933a, b), who examines the effect of stock splits on trading activity, dividend payout rates, and market returns. In recent years, event study models have been growing in popularity. Currie, Kleven, and Zwiers (2020, Figure 4c) summarize trends in working papers from the National Bureau of Economic Research (1980–2018) and papers published in top economics journals (2004–2019). They document a sharply increasing share of papers using event study approaches, with an inflection point around 2012. Typically, event study models are estimated in a reduced-form "treatment effects" context.[1] Applications of event study models vary broadly, from job displacement (as in Figure 1), to school finance reform

[1] They can also be used to estimate statistical moments, which in turn can be used to estimate a structural model, as in Finkelstein et al. (2022).

(Lafortune, Rothstein, and Schanzenbach 2018), to the effect of trade liberalization (Braun and Raddatz 2008).

Behind the scenes of the easily digestible event study picture, a researcher needs to make a number of choices. Some choices are as obvious as the question of how (or when) to deal with pre-existing trends like those shown in Figure 1—and indeed that figure shows different estimates if the pre-event trend is taken into account—and some are more subtle, but researchers are often insufficiently clear about the choices they have made. In this essay, I discuss the range of decisions that go into an event study model, and in this way I aim to improve the understanding of these models for researchers, teachers, and consumers of this research.

For those who wish to dig a few layers deeper, a set of online Appendices provide more detail along with graphic examples and underlying code on related topics, such as connections from event study to difference-in-difference models, showing event study results in a way that is closer to raw data, pooling event study coefficients or using splines over event times to improve efficiency, additional considerations when controlling for pre-event trends, and other topics.

Core Features of Event Study Models

An event study model has two key elements: the estimating equation and the structure of the data.

Estimating Equation

The traditional approach to estimating an event study model is shown in this equation. We have units i and calendar time periods t; in the original example, the units are workers and the time period is calendar time (for example, earnings in the first quarter of calendar year 1982).

$$y_{it} = \underbrace{\left(\sum_{j \in \{-m,\dots,0,\dots,n\}} \gamma_j \cdot D_{i,t-j} \right)}_{\text{Event Study Terms}} + \underbrace{\alpha_i + \delta_t}_{\text{Panel Fixed Effects}} + \underbrace{\beta \cdot X_{it}}_{\text{(Optional) Control Variables}} + \epsilon_{it}.$$

On the left-hand side, the y variable shows the outcome. On the right-hand side, $D_{i,t-j}$ is an indicator variable for event time j, meaning that the event took place j periods before this observation's calendar time. A separate term is included for each event time. The key features of this specification are the $\gamma_j \cdot D_{i,t-j}$ terms. The coefficients after the event has occurred (γ_j for $j \geq 0$) capture the dynamic effects of the treatment as these effects manifest over time since the event. The terms γ_j for before the event has occurred (for $j < 0$) provide a placebo or falsification test. In the absence of anticipation effects, model misspecification, or omitted confounding variables, these pre-event terms should not have a trend in j. Together, this part of the regression equation terms will trace out a graph similar in appearance to Figure 1, measured in event time.

The index t represents the "calendar time" in which we observe the outcomes. The index j represents time-since-event, or "event time." In Jacobson, LaLonde, and Sullivan (1993), event time would be interpreted as, for example, "two quarters after job displacement" (for $j = 2$). In many applications, with at most one event per unit, we can designate the "Event Date" E_i, which is the date that the event occurs. The connection between these three variables is $j = t - E_i$. However, names and labels for the $\{t, j, E_i\}$ variables are not standardized across the literature. As you read event study papers, take care to check your understanding of what names and labels are used for each time concept. The constants m and n determine the endpoints for the estimated event study terms.[2]

Event study models are estimated on data that have a panel structure. It is conventional to add two sets of fixed effects, α_i and δ_t, for unit and time fixed effects. These serve the role of controlling for confounding omitted variables that vary at the unit or time level. Using this two-way fixed events approach helps to isolate the effect of the event. The outcome variable $y_{i,t}$ may also be influenced by other underlying factors. Thus, some event studies add other control variables X_{it}.

Sometimes our events occur at a different level of aggregation than our data. For example, perhaps an event occurs at the state-year level and we are working with a repeated cross section of individual-level data. It is okay to define the event dummies based on the state-year variation and to keep our regressions at the individual level (incorporating cluster- robust standard errors so that inference accounts for the more aggregated level of the event study dummies and their correlation over time within a state). This approach can be useful if we want to control for individual-level covariates; that is, even though we are working with a repeated cross section of individual-level data, we still conceptually have a panel at the state-year level. It can also be okay to first aggregate our data up to the state-year level and then run the model at that level. This makes the dataset more manageable. If we do this, I think it makes sense to weight our aggregated observations by the population represented in each state-year cell in order to get closer to results we would have obtained from the micro data.

Event Study Data Structures

In the panel data used by event studies, units may have an event (in the basic model) or else multiple events (in a more complex model) that occur at certain dates. An event study data structure can be defined based on an understanding of the unit types in your dataset. Two key questions are: (1) Are there "never treated" units or not? (2) Is there (a lot of) variation in the treatment date across units? A researcher needs for the answer to one or both of these questions to be "yes." The combined answers to these two questions represent different data structures, with corresponding differences in the thought experiment behind identification of the treatment effect coefficients. A key theme in this paper is that the options,

[2] In some applications, the time variable t is based on birth cohorts instead of calendar time. This possibility is discussed further later in the paper.

Table 1
Data Structures for Event Study Estimation

	Only Ever-Treated Units	*There are Never-Treated Units*
Common Event Date	N/A	DiD-type
Varying Event Date	Timing-based	Hybrid

Note: Author's proposed labels for event study data structures, based on whether the analysis data sample uses never treated units or not, and on whether treated units have a common event date or varying event dates. "DiD-type" = "Difference in Difference type."

guidance, and conclusions for an event study can depend on the data structure with which we are working.

Table 1 lists the possibilities. In the top-left corner, if we answer both questions with "no" we have only treated units and they share a common event date. In this setting, we cannot separate the effects of the event from other confounders that occur in calendar time, and so cannot identify treatment effects.

If we answer "yes" to the first question and "no" to the second question, then the data include both treated and untreated units, while all treated units share a common event date. The never-treated units help to identify the change in counterfactual outcomes across calendar times. Then the treatment effects can be estimated. In the canonical event studies graph, the treatment effects line represents allowing for over-time changes in the treated group and over-time changes in the untreated group and then looking at the differences between these changes.[3]

In the timing-based data structure, there are only treated units and the event dates vary. A leading example is when different geographic units (or perhaps individuals) all experience the same policy change or treatment, but they experience the change at different (event) dates. Here, the underlying thought experiment is that the timing of the event is as good as random, and so those treated earlier or later can serve as controls for one another. Dobkin et al. (2018) have a timing-based data structure in their study of the effects of hospitalizations on expenditures and labor supply. All of the individuals in their study experience a hospitalization, but they do so at different times.

Sometimes with this data structure, researchers make descriptive event study graphs that omit calendar-time fixed effects and unit fixed effects. For example, Card, Heining, and Kline (2013) track German workers who transition jobs across firms, based on the quartile of wages at the old and new firm. In a

[3] Indeed, the event study specification is a generalization of a standard two-way fixed effects difference-in-difference specification:

$$y_{it} = \gamma \cdot \mathit{Treated}_i \cdot \mathit{Post}_{i,t} + \alpha_i + \delta_t + \beta \cdot X_{it} + E_{it}.$$

Here $\mathit{Treated}_i$ is a binary variable for units that ever receive treatment, and $\mathit{Post}_{i,t}$ is a binary variable that indicates that treatment has occurred. If we restrict the pretreatment coefficients from the earlier equation to be zero, ($\gamma_j = 0$ for $j < 0$), and restrict the post-treatment coefficients to have the same value ($\gamma_j = \gamma$ for $j \geq 0$), then the traditional equation approach shown earlier and this regression here are equivalent.

separate example, Chetty et al. (2014) track Danish workers who transition to jobs with greater defined contribution pension shares. The event study graphs shown in these two examples are essentially expanded pre-post designs. Their credibility comes from three factors: (1) an a priori expectation that the pre-move outcome provides a reasonable counterfactual, (2) the visibly flat pre-trend in the raw data, and (3) the stark jumps at the time of job transition. These graphs have no unit or calendar-time fixed effects and are based on balancing the dataset in event time rather than calendar time.

The data structure might combine variation in event dates and both treated and untreated units. I label this the "hybrid" data structure, and it will include both sources of identification: the comparing of treated and control units and timing-of-event. This data structure is common in event studies. One application that employs a hybrid data structure is the Jacobson, LaLonde, and Sullivan (1993) study mentioned above. They pool data on workers who were displaced at different dates from their jobs as well as workers who were never displaced. Another example is Lafortune, Rothstein, and Schanzenbach (2018), who examine the impact of state-level school finance reforms on funding and test scores. They have 26 states which implement reforms, across a wide range of implementation dates spanning 1990–2011. They also incorporate states without reforms in this period into their analysis.

Estimates from the hybrid data structure can be (informally) compared to estimates relying solely on the timing-based subset of the data (estimated using only ever-treated units) to see whether the different sources of variation are producing similar estimates. So far, I have not seen a formal approach or recipe for making this type of comparison, but I think it could be a useful addition to our standard practice.

When carrying out or interpreting an event study, it is important to be explicit with your reader about the data structure. It is also best practice to show your reader the distribution of observations across event times in your sample. In Appendix A, I place these data structures in the context of related difference-in-difference models. I also illustrate a couple of graphical ways of showing the variation in your unit types and other key aspects of your data structure.

Parameter Restrictions

The basic event studies model includes more parameters to estimate than is possible. Remember, the total number of parameters comes not just from the γ parameters for the treatment effects over each time period, but also from the α and δ fixed effects parameters on units and times and potentially from more parameters if the researcher decides to include additional control variables. More important than a simple count of parameters is the fact that the event-time dummies are multicollinear with the combination of unit (for example, state-level) and time (for example, calendar year) fixed effects.[4] To proceed, we need some restrictions

[4] This multicollinearity is due to the fact that event time, calendar time, and event date are connected by $j = t - E_i$, and that event dates can be defined by unit dummies.

on these parameters. It is useful to group these restrictions into three (potentially overlapping) categories: (1) standard restrictions on panel fixed effects parameters; (2) restrictions that help to define our desired counterfactual; and (3) potential additional restrictions that are required to address concerns about multicollinearity.

In an event study model, the event-time coefficients γ_j in the traditional equation approach shown earlier are our main coefficients of interest. They estimate the treatment impact *j* periods after receiving treatment. This treatment impact needs to be defined in reference to a specific counterfactual. That definition is embodied in parameter restrictions. For example, we might think of a difference-in-difference-type counterfactual as "compared to a pretreatment period, how much change we would have expected to have occurred in the absence of treatment." Thus, probably the most common normalization is to choose a specific pretreatment event time and normalize the corresponding coefficients to average to zero. For example it is a common choice to set $\gamma_{-1} = 0$, by excluding the dummy variable for the $j = -1$ event time from the regression. Alternatively, we might have experimental assignment to treatment and control unit types. In this case, our normalizing assumption might be that those in the untreated group can serve as a control group for those who are treated. We would, therefore, have all of the event-time dummy variables but omit the unit fixed effects, setting $\alpha_i = 0$.

Multicollinearities abound in event study models. At a basic level, the sum of the unit dummies is equal to one, and the sum of the calendar-time dummies is equal to one. This introduces a multicollinearity between these two sets of dummies as well as the intercept, typically requiring dropping one from each set. There is also an additional multicollinearity between the event-time dummies $D_{i,t-j}$ and the unit and calendar-time dummies. Sometimes, once we have made basic restrictions on fixed effects and to define the counterfactual, the remaining parameters in the model can be identified and we are good to go. But this is not always the case. The problem of multicollinearity is especially prevalent in a "timing-based" data structure, where all units are treated but their event date E_i varies. In this data structure, the event-time dummy variables, unit dummy variables, and calendar-time dummy variables will have one or more additional multicollinearities, and so additional restrictions are needed in order to proceed.[5] The problems of multicollinearity also compound when we directly add in unit-specific time trend controls.

How should we implement our additional required parameter restrictions? In current practice, a common approach is to let the software (like Stata) automatically choose some collinear variables to drop, with unknown and possibly problematic implications. This approach should be avoided, and my recommendation is to check your regression output carefully to ensure that no variables are being unexpectedly dropped.

[5] See Proposition 1 in Borusyak, Jaravel, and Spiess (2022) and section 2.4.2 of Schmidheiny and Siegloch (2023). I discuss the number of needed parameter restrictions further in online Appendix B.1 and different examples are illustrated in online Appendix C.2.

Another common approach is to pool some of the data by grouping several of the treatment variable γ's in the tails to be equal. In the traditional event study equation, this would mean including an "end-cap" dummy variable such as $D_{i,t \leq E_i - m}$, indicating "the event will happen m or more periods in the future." This approach can sometimes be okay, but it can be problematic if there are uncontrolled-for underlying trends or (for a posttreatment end cap) if the treatment effects themselves are trending. It should only be used if these concerns seem unlikely to be important. As an alternative, it is possible to apply milder but still-useful constraints. For example, you can focus your parameter restrictions on the pre-event coefficients. I discuss "end caps" more in the next section.[6]

It's not always obvious when our model is okay as is or when additional restrictions are needed. When researchers need to impose additional restrictions to identify the model, we should keep in mind the following: (1) these are not merely formalities—the treatment effect coefficients γ_j we estimate are directly dependent on the restrictions imposed; (2) these restrictions are untestable, at least in part; and so (3) we want for these to be as uncontroversial and "obviously true" as possible. Indeed, (4) because of the "multicollinearities abound" nature of some event study data structures, our main estimates of interest can be unexpectedly sensitive to these extra restrictions. This can result in (5) "small bits of noise" propagating through our model in unexpected ways. This last fact can sometimes argue for employing additional restrictions beyond what would be minimally necessary.

My main recommendation is to be clear and explicit about what restrictions are being imposed. Going forward, it would be useful if all event studies would clearly report (1) the number of categories for each group (time, unit-type, or unit) of dummies and/or event study coefficients, both the total possible as well as those that are included in our actually estimated specification (after variables are dropped due to collinearity); (2) the constraints we (or our statistical package) impose on the estimation, either directly or through dropped terms; and (3) a direct assessment of the identifying variation in your data structure (for example, by computing the rank of the relevant proportion of your X matrix).

Event Study Specification Choices

This section outlines some of the main specification choices to be made when estimating event study models and discusses the trade-offs involved.

Choice of Pre-event Reference Period

When estimating an event study model, a common choice is to use "one period before treatment" as a normalization, so that the γ_{-1} coefficient is set equal to zero

[6] In online Appendix B.2, I offer more detail and examples for parameter restrictions, including some discussion of useful Stata commands.

in the time period immediately before the event. In the traditional event study equation presented earlier, this is implemented by dropping the -1 event-time dummy variable. But instead of blindly choosing the period immediately before the event for the normalization, it is better practice to make a judgment call as to what is a reasonable pre-event window, balancing considerations of "close enough to be the appropriate counterfactual baseline" and "more data allows for more precision." Then all of the event dummies can be included and the γ_j coefficients constrained to average to zero within the pre-period window.

How long of a pre-event window should a researcher choose? There is no hard-and-fast rule. I think it is useful to consider the pre-event window you would choose if you were estimating a simple difference-in-difference model. If you chose just one pre-event-time period, you might be worried about the extra statistical noise this would bring. As your pre-event window extends farther back, at some point you might get increasingly worried that those time periods become less appropriate for your counterfactual. In the end, for your difference-in-difference model you would make a judgment call, trading off these two considerations. It seems sensible to have this same judgement call inform your choice of the pre-event reference period.

Normalizing to zero over several event times, rather than just the period immediately before the event, has two effects on the canonical event studies graph. Choosing a longer time period has the effect of shifting the whole pattern of coefficients up or down—while retaining the same shape. The other effect is that when a more extended reference period is used for normalizing to zero, the standard errors can be noticeably smaller. The reason is that when using a single reference time period there is additional uncertainty driven by the noise in this term on its own, which tends to make the standard errors larger.[7]

If we normalize to a broader reference period, our search for a trend before the event will manifest itself differently than if we had normalized the –1 coefficient to zero. We need to assess the overall trend in coefficients rather than examine pointwise coefficients and their difference from zero. (This is also illustrated in online Appendix C.1.)

When we suspect (or see) a dip in outcomes shortly before the event, we might speculate that this is driven by some process that is bundled with the event and which is playing out shortly before the event as it is recognized in our dataset. In this case, we probably do not want to use the period of the dip as our counterfactual baseline because it is actually part of the treated period, even though nominally it's before treatment.[8] Instead we could define our baseline counterfactual to be a period prior to the beginning of the dip. For example, in Figure 1 we see a dip in

[7] This is illustrated in online Appendix C.1. Also, online Appendix C.2 illustrates the potential impact of different normalizations within a timing-based data structure.

[8] A dip that occurs just before the event is sometimes called "Ashenfelter's dip," after Ashenfelter (1978), who studied the impact of job training on earnings. Ashenfelter's models were not presented in the now-traditional event study graphical format, but his table's results have an event study framing, including showing a pretraining drop in earnings.

earnings prior to the layoff. Inspection of the figure suggests that we would want to have our reference period be at least one year prior to the layoff. In Jacobson, LaLonde, and Sullivan (1993), the authors chose "5 or more years prior to the layoff" as the reference period.

Show More than the Estimated Treatment Effects

An event study provides a treatment estimate as a single set of numbers. However, it is good practice to get closer to the raw data by also reporting a combination of actual and counterfactual average outcomes separately for each unit type. These graphs will complement each other in terms of the information provided. For example, when the event study allows comparison of treated and untreated units, this presentation allows readers to assess whether the unit types experience parallel trends during periods when treatment status is unchanging. Both the difference in levels and in the trends can provide important context for interpreting the treatment effects.

We can also add to this plot a line for the counterfactual untreated prediction that applies to the treated units. To generate this, here are the appropriate steps: (1) estimate the event study model; (2) "zero out" the event-time dummies and make predictions; (3) average these predictions within calendar time for the treated units; and (4) plot out this counterfactual. This calculation lets us see both the raw data and the estimated treatment effects. It also implicitly shows the content of the normalizing restrictions of the model. For example, if we are normalizing the pre-trend in event studies coefficients to be zero and are controlling for unit-type trends, this will show up in a trending counterfactual line. For timing-based or hybrid data structures, this lesson is slightly more complicated to apply. However, the researcher can still plot the average time series for each unit type and then supplement this by adding the counterfactuals for each unit type. (These ideas are illustrated in Appendix D.)

Choices with Control Units: Selection and Re-weighting

Suppose that we are carrying out an event study that includes both treated and untreated units (for example, individuals or states), with untreated units as the control group. However, sometimes we might worry that the never-treated units could be problematic comparisons for the treated units. For example, Krolikowski (2018) reconsiders the Jacobson, LaLonde, and Sullivan (1993) example that generated Figure 1 presented earlier. In the 1993 paper, the event is "first observed layoff"; never-treated units are therefore individuals who never experienced a layoff. However, subsequent layoffs can only occur for the treated group. Thus, there is a mechanical difference in the future earnings potential of the treated group compared to the control observations, above and beyond the effect of the first layoff under consideration. In addition, the control group may be positively selected with regard to unobservable skill, labor force attachment, and/or job match quality. In this setting, those who are never laid off may not provide a good counterfactual for outcome for treated individuals; indeed, the use of this control group could make the impacts of the layoff look worse than they actually are.

There are a range of options to have the control units (for example, individuals who did not experience job displacement) offer better counterfactuals, with the overall goal of making the assumption that "the control units tell us the counterfactual over-time changes" more plausible. First, a researcher might exclude a subset of the control units because they are in some way unrepresentative or because they experienced unusual shocks. For example, if you are working with a city-year panel, and your treated cities are all medium- or large-sized, then you might consider excluding small cities from the control units that you use. Second, for the time periods before the event, it is possible to check for parallel trends between the control and treated units. Third, one can look at the degree of similarity between treated and control units along a number of dimensions, using covariates.

Finally, the researcher might use a reweighting or matching procedure prior to estimation of the event study. A reweighting procedure would apply different weights to the never-treated units so that the covariates match the treated units. In a study of the impact of the introduction of the Legal Services Program (during the 1960s) on demographic outcomes, Goodman-Bacon and Cunningham (2019) observe that untreated counties are different in their observables compared to treated counties. To address this, they estimate a cross-county first stage model to obtain propensity scores (specifically, the probability of being a treated county). They then re-weight the control counties to be more representative of those treated.[9]

An alternative is to choose one or more never-treated "matches" for each treated unit. These matches would typically be made based on observable covariates, possibly including some values of pre-event outcomes. Some practitioners choose to combine these approaches with assigning a pseudo-event time to each control unit, in an effort to present a more plausible counterfactual outcome path. I am not aware of a systematic look at possible trade-offs involved in the choice to use pseudo-event times for the control units.

If an event study has a hybrid data structure, a researcher has the option of discarding the data from untreated units and focusing instead on a timing-based strategy. This approach that would be based on the belief that "among those treated, timing of treatment is as good as random" is more believable than the assumption that "control units tell us the counterfactual over-time changes." On the other side, using never-treated units will bring in more data, usually improving statistical power and requiring fewer parameter restrictions in order to identify the model. The trade-off between these two considerations will vary on a case-by-case basis. Whatever approach is chosen for dealing with never-treated units, it is useful to show sensitivity of the results to alternate approaches.

[9] This approach could in principle also be used in situations that use only ever-treated units. If there is reason for concern over possible differences between, say, earlier-treated and later-treated events, We could use reweighting to balance covariates across "early event date" and "late event date" units prior to estimating the model.

Choice of Event Window

In some cases, data availability will limit what endpoints *m* and *n* can be used for the event-time window; otherwise, you need to make an explicit decision. On one hand, making the event window as wide as possible allows us to see a long path of dynamic treatment effects, and for the pre-event coefficients it gives us a long window to detect troublesome patterns. This consideration pushes toward including as many event-time lags as possible.

The main competing consideration is that we would ideally like for the event-time coefficients γ_j to all be estimated off of the same set of units. For example, in Jacobson, LaLonde, and Sullivan (1993) the events (job displacements) occur between 1980 and 1986, and the outcome (earnings) data are observed for the period 1974–1986. The event-time coefficients for "zero years since displacement" in Figure 1 are based off of all displacements. But the coefficients for "five years since displacement" can only be estimated for displacements that occur in 1980 or 1981. This means that the event-time coefficients post-displacement are estimated off of different sets of individuals. If there is something systematically different about the early- or late-displacement individuals, that could challenge interpretation of the coefficients. Even without a systematic difference, there will be a loss of statistical power as fewer units are available to identify the more remote coefficients further from the event itself. These considerations suggest if possible choosing the endpoints of the event window so that most or all coefficients are identified off of a balanced set of units. It also reinforces the need to show your reader the distribution of data across event times (as illustrated in Appendix A).

Depending on your data setup, there may be a straightforward resolution of these competing concerns. Suppose that the span of event dates lies within a ten-year window and that you have data for at least 20 years on either side of that window. Then it might be easy to focus on event-time endpoints that are within 20 years and have a fully balanced set of units for each event-time coefficient. But even in this case, if you observed that the interesting dynamics in terms of treatment effects are resolved within the first five years of treatment, it might make sense to limit the event window to eight to ten years, to bring more visual attention to the period of interest and show the leveling off.

If your data setup does not allow for a straightforward resolution, then you need to make a judgment call. In this case, it will be useful to offer a "robustness check" specification, in which you choose an alternate approach (such as a wider event window).

Finally, it will be important for readers of an event study to know the degree of balance or imbalance in the number of units available to identify the event coefficients. This could be discussed in the text or presented as an appendix table showing the count of units, by event time *j*.

Special Attention for the Endpoints?

In event studies, it will be common to have data for some units that occur before or after the event window. In the notation of the traditional event study equation, these would be observations for which $j \leq -m$ or $j \geq n$. We need to decide how to address this issue.

One natural option is to create and include as many event dummies as possible. By directly estimating a γ_j for each event time, this removes the problem. This approach is natural and appropriate when the data structure has both treated and untreated units that are balanced in calendar time (for example, all US states are observed over the period 1980–2020).

A second option involves creating "end-cap" variables in the traditional event study equation. For example, the data before and including the "pre" endpoint might be given a common dummy variable, $D_{i,t \leq E_i - m}$. Similarly the data points after the "post" endpoint can share a common dummy variable. I think this choice is the most common one, and often it is a good one. Schmidheiny and Siegloch (2023) recommend this approach (which they call "binning"). They note that it can provide a natural identifying restriction for timing-based data structures, that it creates a natural connection to distributed lag models, and that it can lead to a straightforward way to model multiple events per unit.

The main risks to creating "end caps" arise with trending counterfactuals or trending treatment effects. These risks are discussed further in the next main section of the paper on trends. That section argues that we might be hesitant about including "post" end caps if we think that there is a chance that treatment effects are changing over event time.

Another possible approach is just to drop observations that have event dates outside of our main window of interest. This option keeps the specification simple and creates a balance in event time in our analysis sample (for example, all US states are observed from three periods before their event to five periods after). One possible trade-off is that the loss of data can weaken statistical power. An additional consideration arises when using only ever-treated units: with this data structure you can be balanced in calendar time or balanced in event time, but not both. If you limit your sample to be balanced in event time, then this creates an imbalance in calendar time. This in turn means that the time dummies at the extremes will be estimated off of strangely selected units. Because the time dummies play a fundamental role in the identification of treatment effects, this approach seems risky to me.

A final option is not to include an event-time variable that is turned on for these faraway observations. In this way, the faraway observations are pooled together as part of the reference group for when the event did not happen. For example, the reference group in Figure 1 appears to be "more than five years before job displacement." This choice can be acceptable, but you should not include both "before the first endpoint" ($j \leq j_{min}$) as well as "after the final endpoint" ($j \geq j_{max}$) in the same reference group. Also, you should not combine "before the first endpoint" with the time period before the event in the same reference group. For example, in Figure 1 event time –1 is not part of the reference group.

A related choice when presenting a graph of the event-time coefficients concerns whether and how to plot endpoint coefficients. When the endpoint has its own dummy variable, it will capture different averaging than the "interior" terms and will sometimes appear to be offset a bit from the rest of the graph. This can distract the reader from the main story about what is going on closer to event time zero. On the other hand, including such endpoints in the graph gives a fuller picture of the model; indeed, including them can sometimes help to diagnose problems with the specification or the data. I think best practice should typically be to plot the endpoint coefficients and to indicate in the figure (whether with a distinct symbol and/or in the figure notes) that these are differently estimated from the other event study coefficients.

Overall, you need to explicitly decide how you will deal with the endpoints and inform your readers about your decision.

Pooling Event Times for Statistical Power

With so many "key coefficients" to estimate, event study specifications can ask a lot of the data. Many event study models have pretty wide confidence intervals around each of the main γ_j coefficients. One strategy to regain some statistical power is to estimate models that pool together two or more adjacent event-time dummies, and then include these pooled variables in the model instead of the single-year event-time dummies. This approach strives for a balance between flexibility and statistical power. The main risk is that the pooling might obscure features of the empirical results. If you do this pooling, it is probably best to also show results from the unpooled model as a robustness check.

There are a variety of ways to pool event-time data. For example, one can restrict the model so that the coefficients will be the same in, say, periods 1 and 2, periods 3 and 4, and so on. Goodman-Bacon (2018) uses pooled event-time dummies to present results in table format. A more complex alternative is to restrict the event study coefficients to lie on a spline function between the points—essentially forcing a kind of averaging across points, but allowing for a flexible functional form (in a piecewise linear spline, the event study coefficients are forced to lie on a connected set of straight lines). For example, Bailey et al. (2020) and Lafortune, Rothstein, and Schanzenbach (2018) use spline restrictions for improved statistical power. However, this approach comes with some risk of mischaracterizing the pattern of treatment effects, in particular if the imposed model is not flexible enough to reflect reality. When using splines, it can make sense to allow for a jump or break in the splines in the transition from pretreatment to posttreatment periods. Lafortune, Rothstein, and Schanzenbach (2018) implement a model with a linear trend in event time, a jump at event time 0, and then a separate linear trend for event times after the event. As with pooling, it is best practice to also show the unconstrained model as a robustness check. Appendix E offers examples of pooling coefficients and spline models.

The Problem of Trends in Event Studies

Trends can cause problems for event studies in two distinct ways. First, treated unit types might follow a different trend than untreated types in terms of their untreated potential (and unobserved) outcomes, which can confound the estimated treatment effects. As illustrated by Figure 1 at the start of the paper, if a trend is already apparent before the event, it calls into question how to interpret patterns after the event. Second, treatment effects themselves may be trending in time-since-treatment. This second possibility is not necessarily a problem for event study models: after all, the point of these models is to allow for treatment effects that vary over time. But trending treatment effects can cause problems for the estimates from certain specification choices. In this section, I lay out these issues and some possible approaches in more detail. (Appendix F has an expanded discussion and graphical illustrations for several of the main points.)

Pre-event Coefficients as a Diagnostic Tool

The estimated pre-event terms can serve as a tool for diagnosing trends. This is often done informally by inspecting the graph of the pre-event coefficients. This tool is most appropriate when working with difference-in-difference or hybrid data structures, which include never-treated units.

An additional consideration arises if we are working with a timing-based data structure with no control units. In this setting, Borusyak, Jaravel, and Spiess (2022) show that due to the multicollinearity of event-time, calendar-time, and unit fixed effects it is impossible to identify a linear trend in the set of treatment effects (or in the pretreatment coefficients). In this case, the best we can do is to look for nonlinear pre-trends. For this data structure, they recommend the normalization of setting an additional pretreatment γ_{-a} coefficient to be zero. This step imposes a zero pre-trend, and allows for visual or statistical inspection of the other pretreatment coefficients as a check for nonlinear pre-trends. Schmidheiny and Siegloch (2023) argue that using end caps can provide identification of the event-time coefficients in a timing-based data structure. This would restore the ability to examine pre-event trends.

A separate difficulty is that if you have too few pretreatment periods, it can be hard to distinguish between actual pre-trends and statistical noise. This limits the comfort a researcher can take from "passing" a test of no visible pre-trend. There is no hard and fast rule for "how few is too few." When looking at a graph of event study coefficient estimates, I find it useful to mentally visualize the range of possible pre-trends that could be consistent with the pretreatment estimates. Across papers that I see, this approach often leaves me feeling skeptical if I see three or fewer pre-event terms. But this depends on both the variability and the apparent trends among those pretreatment coefficients. If you are concerned about this issue, a simple additional step here is to add more pretreatment periods, extending further back in event time. In a more structured approach, Dobkin et al. (2018) plot the linear pre-trend from a parametric model on the figures that show event study coefficients.

Recent econometric work identifies some potential problems with using pre-trends as a diagnostic tool. Roth (2022) notes that the widespread, informal practice of "rounding insignificant pre-trends to zero" can lead to "pre-test bias."

Even a mild pre-trend, which cannot be visually or statistically detected, can still meaningfully influence the estimated posttreatment impacts. Roth argues that if we are confident of the functional form of the trends (for example, that the trends are linear in time) we should plan always to control for trends regardless of whether or not there is not a strongly apparent pre-trend. He also presents more sophisticated extensions to methods of controlling for trends, based on Freyaldenhoven, Hansen, and Shapiro (2019) and Rambachan and Roth (2023), that allow a researcher to proceed under weaker assumptions about the functional form of the trends.

Separately, pre-trends can be biased if our underlying model is misspecified. Sun and Abraham (2021) examine the case where the misspecification arises from different unit types having different treatment effects—say, if those treated earlier in calendar time have larger treatment effects than those treated later in time. For example, suppose that in Jacobson, LaLonde, and Sullivan (1993) the individuals facing job displacement early in the sample (1980–1982) have greater impacts than those displaced later in the sample (1983–1986), perhaps due to changes in the macroeconomic environment. This difference in treatment effects can lead to a (spurious) apparent trend in the estimated pre-event coefficients. In this case, the appearance of a pre-trend is an indication that something is wrong with the specification of our model.[10] De Chaisemartin and D'Haultfœuille (2022) propose alternative pre-trend estimators that are robust to different unit types having different treatment effects.

Controlling for Unit-Specific Trends

Rather than focusing on pre-trends as a diagnostic measure, an alternative is to control for unit-specific trends, by including a (continuous) time variable interacted with unit (for example, state) dummies. This approach is suitable if we believe that pre-trends reflect trending omitted variables that could bias the main estimates of the treatment. Controlling for unit-specific trends aims to eliminate this omitted variables bias.

For example, Alsan and Goldin (2019) use an event study to examine the historical introduction of clean water and sewer projects across municipalities during 1880–1920. In their specification they control for municipality-specific time trends. In a separate example, Bostwick, Fischer, and Lang (2022) study the impacts of a university switching from a quarter-based to a semester-based schedule. They want to make sure their estimates are not confounded by outcomes trending differently across universities, so they control for university-specific time trends.

What are the main trade-offs between using pre-trends as a diagnostic and controlling for unit-specific trends? First, controlling for unit trends changes the counterfactual thought experiment. Our treatment effect estimates now have an

[10] The specification error here is the assumption of treatment effects that are the same for both early and late treated units. This contrasts with our usual interpretation of pre-trends as indicating anticipation effects or different underlying trends in untreated potential outcomes.

interpretation of "my outcome compared to the reference period, and net of underlying linear trends in the counterfactual between that period and now." Second, because the counterfactual now controls for these trends, our pre-trends should look flat by construction. Thus, we lose the basic falsification test that the pre-trends provide in the basic model. However, one can still use the pre-event coefficients to look for nonlinear violations of the parallel trends assumption.

Adding unit trend controls may also interact uncomfortably with the "too many variables" and "multicollinearities abound" challenges of event studies. A unit-specific time trend term will be multicollinear with the unit dummies, time dummies, and event-time variables. This means that it will be necessary to add (at least) one more additional parameter restriction. As noted earlier, it is imprudent to just let our software address the collinearity problem by dropping a variable on its own, because what it drops might undermine our interpretation of the resulting estimates. The choice of which restriction to apply is guided by the same principle as before: the additional restrictions are untestable, but estimated coefficients on the trend terms will be affected by the restrictions we place, so we want it to be as "obviously true" as possible. Our additional restriction should be applied to the event study coefficients γ_j so that we do not undermine the panel fixed effects controls in the estimation. A common choice is to impose a restriction that two event coefficients are equal. The trend estimates will be estimated in the context of those restrictions; in particular, the later treatment estimates may "pivot" as a result of imposing this restriction. (This is illustrated in online Appendix F.)

Trends versus Pre-trends

The estimated parameters for unit-specific trends will seek to capture trending behavior both before and after the event. But what if treatment effects are also trending? Suppose that in our traditional event study equation, treatment effects are increasing in event time: $\gamma_0 < \gamma_1 < \gamma_2 \ldots$. Then, an estimate of unit-specific trends might try to fit both pre-event trends *and* the treatment effect pattern. This in turn can bias the estimated event study coefficients. This problem can occur if our parameter restrictions include post-event terms, such as a post-event end cap. The post-event end cap (for example, "six or more periods after the event") forces all the estimated event-time effects γ_j within the end cap to be the same. If instead they are truly trending, this can cause problems. An extreme version of this is a difference-in-difference specification, which restricts all post-event treatment effects to be the same.

This is the main argument in the Wolfers (2006) critique of prior difference-in-difference approaches examining the impact of unilateral divorce laws on divorce rates. Much of this work used state trend controls. Wolfers argues that these laws will have dynamic impacts—that is, trending γ_j for $j \geq 0$. Because of this, the estimated state-specific time trends will be contaminated by trying to also fit the trending treatment impacts. This in turn will bias the main estimates. Wolfers proposes as an improvement a variation of an event study specification for the post-event periods. (I present a stylized illustration of this phenomenon in online Appendix G.2.)

One option to prevent this problem is to focus on controlling for "pre-trends" only. However, this may require custom programming (online Appendix F.4 presents one approach). Another option is to model your event study terms (the $\gamma_j \cdot D_{i,t-j}$ terms in the traditional event study equation) so that the unit-specific time trends will not be confounded by trending treatment effects. For example, one can drop all constraints on the event study parameters for posttreatment by including all "post" event-time dummies and having no "post" end cap. This step will ensure that the trend coefficients are estimated based only on pre-event data.

Recommendations in Controlling for Trends

These considerations lead to guidelines for researchers who are controlling for trends. First, don't let post-event parameter restrictions influence your estimated trends, unless you are highly confident that the treatment effects are not trending in that range. Otherwise, your control for trends may be picking up part of the trending treatment event.

Second, if one of your extra parameter restrictions is in the form of equality of two event coefficients, consider spacing those coefficients further apart, because the impact of any statistical "noise" between the two coefficients will be larger if they are closer together. As an alternative, focus the restrictions on event study parameters that allow for more averaging across units.

One restriction that accommodates the considerations above—at least for the difference-in-differences and the hybrid data structures—is to constrain the trend in the "reference period" event study coefficients to be zero. This approach has the advantage of averaging across several coefficients, and reducing the impact of noise from any one or two of them. It also respects the notion of having a reference period embodied in the normalizing restriction (as discussed earlier), and offers a natural counterfactual interpretation: "Compared to the level and trend in the reference period, and the over-calendar-time changes from control units, what would my expected outcome be?"[11]

Finally, if we are working with a timing-based data structure, controlling for trends has potential to create surprising and severe problems. Adding *linear* trend controls (and the required additional parameter restriction) can induce *quadratic* trends into our estimated event study coefficients. This arises from a subtle way in which the event dummies are collinear with the other variables in the model (as illustrated in online Appendix F.2). The key lesson is to be extra cautious about the combination of trend controls and a timing-based data structure.

There is one way in which adjustments for trends can often be simplified compared to common practice: we can focus on unit-type trends, rather than

[11] Given a reference period (k_1, k_2), this is implemented with the following linear restriction on the event study parameters: $\sum_{j=k_1}^{k_2} \left(j - \frac{k_1 + k_2}{2}\right) \cdot \gamma_j = 0$. To derive this, consider a bivariate regression $\gamma_j = \phi_0 + \phi_1 \cdot j$. To impose a zero trend, we want $\hat{\phi}_1 = 0$. The left hand side of the proposed restriction is the numerator for the coefficient $\hat{\phi}_1$.

unit-specific trends. That is, we can allow for one shared trend parameter for each group of units that share an event date. This is because the event study variables depend only on unit type and time. Once we condition on unit-type trends, any remaining unit-specific trends will be orthogonal to the event study dummies and will not influence their coefficients.

Statistical Inference for Event Study Models

For researchers, the usual primary concern is to have unbiased point estimates. However, we also need to be able to conduct statistical inference to test hypotheses about the true state of the world.

Cluster-Robust Inference

For event study models, the current practice appears to be to calculate cluster-robust standard errors, with clusters defined as the *i*-level units. This starting place is sensible. The key right-hand-side variables in an event study have some degree of autocorrelation and it is plausible to think about the model error term also being positively autocorrelated over time within a unit. Taken together, this argues for clustering at the unit level. The general rule of thumb is that we want to cluster at a level when there is correlation in the scores ($X_i e_i$, driven by correlation in the model errors e) across units in that cluster (Cameron and Miller 2015, section II.C). If the underlying event is shared across units, then this argues for clustering at a higher level. For example, if our dataset is a panel of individuals, but the event is a state-level policy change, then we likely want to cluster at the level of the state.

However, one potential concern is that standard cluster-robust methods provide accurate standard errors only if the number of clusters is "large enough," with no hard and fast rule for what that means. Folk wisdom and some simulations offer rules-of-thumb like 42 or 50 clusters, but in some settings this is not enough, and in other settings a smaller number will suffice. When there are too few clusters, traditional cluster-robust methods may over-reject. If we are facing too few clusters, we need to take account of this in our inference procedures (Cameron and Miller 2015, section VI).

The problem of few clusters is exacerbated when the clusters are of asymmetric size or when there are very few treated units. In these settings, our conclusions need to be more tentative. But it is not so bad that you just have to give up. In these settings, the adjustments offered in Imbens and Kolesar (2016), Carter, Schnepel, and Steigerwald (2017), and MacKinnon and Webb (2017) might be a good choice.[12]

[12] Permutation tests are an alternative approach to conducting inference. The idea is to randomly reassign pseudo-event dates across units, and re-estimate the model. Repeat this procedure many times, to construct a distribution of "estimated treatment effects, when there is no actual treatment." A distribution of test statistics can be constructed from these permutations. The main estimates can be compared

The Spatial Correlation Problem

The basic premise of cluster-robust inference requires that clusters are independent from one another. Spatial correlation in event dates undercuts this premise, and doing so may result in over-rejection of the null hypothesis. When events are the result of a political process or influenced by economic circumstances, neighboring units (say, neighboring states) can be closer in event date than more distant units. Often, economic outcomes are also spatially correlated.

For the most part, the current empirical literature ignores this concern, but there are some potential ways to address it; for example, see Conley (1999) on spatial robust standard errors. However, there is little guidance on how to on measure "distance" across some combination of space and time.

Another possibility is to allow for arbitrary correlations in observations within a cluster, and also allow for correlation that decays in calendar time across observations in nearby time periods, regardless of the unit to which they belong (as in Driscoll and Kraay 1998). This approach allows for greater dependence across observations than the current standard. But there is no "button to push" for implementation of these approaches, so it will require custom programming. In addition, allowing for spatial autocorrelation is likely to make the "few clusters" problem even more salient.

For the near term, a basic precaution is to examine your data for the possibility of spatial correlation in event dates, although currently there is no hard-and-fast guidance for what levels of spatial autocorrelation should be a matter of concern. For now, the standard practice of "cluster on the underlying event" seems likely to continue. However, researchers should probably start to pay more attention to spatial correlation in the future.

Extensions and Challenges

My discussion has focused on a basic version of event study models. In this section I briefly note a few of the additional extensions and challenges that may arise.

Events with Variable Intensity

What should researchers do when events can vary in their magnitude? For example, suppose the event is a cigarette tax hike or an increase in the state minimum wage. We might want to allow for the event to scale proportionally to the size of the shock. This issue can be handled in a straightforward way by pre-multiplying the

against these distributions, and if they fall in the tails of the distribution, this is evidence against the null hypothesis of no impact. MacKinnon and Webb (2019) study this randomization approach in a difference-in-difference setting with few units, very few treated units, and clusters having different sizes (for example, larger and smaller US states). Young (2019) also shows that randomization procedures (based on t-statistics) perform well. I think it likely that such results would carry over to the event study setting.

event dummy by the magnitude of the event. For example, the event variable $D_{i,s}$ could be "by what percentage were cigarette taxes hiked?"

One interesting variation is found in Goodman-Bacon (2018), who examines the introduction of Medicaid in the 1960s across US states. The impact of introduction varied state-by-state as a function of the fraction of population that was receiving assistance from the Aid to Families with Dependent Children welfare program at the time when Medicaid was introduced. This setup combines both timing-based and variable-intensity variation in treatment. Concerned about event dates being correlated with preexisting trends, Goodman-Bacon (2018) isolates the variation from variable-intensity of treatment from the timing by including dummies to control for event date by calendar year.

A group of recent papers center event study models within "mover" strategies. These include consumers changing purchase patterns as they move across locations (Bronnenberg, Dubé, and Gentzkow 2012) and either doctors (Molitor 2018) or patients (Finkelstein, Gentzkow, and Williams 2016) moving from one region to another with different patterns of health-care practice. These mover designs often pair with variable intensity of treatment. For example, Finkelstein, Gentzkow, and Williams (2016) track Medicare patients who move across regions with different intensities of medical usage. In this case the event is the move, and the variable intensity reflects the difference in medical usage between the destination and origin locations. Molitor (2018) examines cardiologists' patterns of practice as they move across regions. Again, the variable intensity of the event is given by the difference in regional patterns of practice across destination and origin locations.

More than One Event Per Unit

What if there is a possibility of multiple events per unit? For example, the data might include repeated layoffs or repeated state minimum wage hikes.

For the basic case, this can be straightforward to implement. We define the event $D_{i,s}$ to be one in any period where an event occurs, and we allow this to happen in different time periods for the same unit *i*. Thus, more than one of the event-time dummies can be turned on simultaneously. Sandler and Sandler (2014) suggest this approach.[13] However, this approach provides a specific interpretation of the estimated coefficients—a "partial effects" interpretation, holding constant the potential impact of subsequent events (including those whose existence might in turn be impacted by the current event). This can be different from the "total effect," which includes the impact of today's event on the likelihood of future events happening. Krolikowski (2018) explores this issue by using a simulation to

[13] A related but alternative approach is to duplicate data around the event. In this approach, each observation is "split" into multiple new observations based on unique combination event-by-underlying-unit. However, Sandler and Sandler (2014) show using Monte Carlo simulations that in some settings this can lead to biased estimates.

propose a weighted average of the partial effect estimates as well as a "first event only" model.[14]

Another approach is to adjust the definition of an event so as to have only one per unit. For example, in the Jacobson, LaLonde, and Sullivan (1993) paper behind Figure 1, the focus is on the first layoff, and subsequent layoffs are not modeled. This approach could be implemented based on "biggest event" or "first big event." Again, issues will arise in interpreting the resulting coefficient. By bundling subsequent events (and their dynamic impacts) into the definition of "treatment," we have a potentially nonintuitive definition of treatment—a version of the partial-versus-total effects problem just mentioned. In some cases, this approach is combined with using the "never treated" group as a control group. This combination can introduce the selection issue mentioned earlier—that is, the control group may now differ in unobserved ways, like stronger skills or labor market attachment, so comparisons with them will give biased counterfactuals for the treated. This can make the estimated effects of the displacement look worse than the true causal effects. Separately, it can raise concerns about external validity of the findings to the broader population.

The possibility of multiple events raises the question of whether the effect of an event depends on the history of other events. A first layoff is one thing, but we can imagine that subsequent layoffs are possibly worse (increasing fragility) or not quite as bad (either toughening up or "floor effects") as the first. In principle, the basic model could be modified to estimate sensitivity in treatment effects directly, based on the history of prior events, but I have not yet seen this implemented.

Heterogeneous Treatment Effects

What if the effect of treatment does not just vary in "time since event," but also depends systematically on the unit type, time, or context? For example, the treatment effect might depend on observable variables or on the date of adoption of the event.

Suppose we are interested in how a treatment effect varies across men and women. We can then include one set of event dummies for men and another set of event dummies for women. More generally, we can include a set of interaction terms, based on the covariates that we believe influence the treatment effects. In taking this step, it is important to follow usual best practice for interaction terms in regression models, such as including direct controls for the covariates if they are time-varying. These interactions can use up a lot of variation in the data, and in response, it may be useful to impose a parametric simplification on the interaction terms. For example, the Jacobson, LaLonde, and Sullivan (1993) example

[14] Basso, Miller, and Schaller (2022) label the partial effect "Y channel only (YCO)" and contrast the Event Study approach with a Local Projections approach to estimating dynamic treatment effects. They observe that Local Projections can directly recover the "total effect" (corresponding to the impulse response) and show that it can be transformed into the YCO. Cellini, Ferreira, and Rothstein (2010) also address this distinction in the context of a dynamic regression discontinuity model and estimate both effects. They label the partial effect "Treatment on the Treated" and the total effect "Intent to Treat."

from Figure 1 used a parsimonious approach of having three periods of treatment effect: the "dip" (the 13 quarters prior to job displacement), the "drop" (the quarter of displacement), and the "recovery" (six quarters following displacement). They allow covariates to produce different slopes (in event time) for these three periods.

In a setting with varying event dates, we might want to model the possibility that the dynamic treatment effects depend on the timing of adoption. For example, US states that are early to adopt a policy might be the ones that benefit the most from that policy; late adopters might have less or even opposite-signed effects. One approach is to treat the actual event date as an observable variable and estimate treatment effects based on the date, or perhaps using an "early"/"late" adopter dummy variable. This approach should work, so long as there are control units or enough variation in event dates.

There is a recent, active, and promising literature on how event study models perform when treatment effects differ across units and when we do not know the functional form of how they differ. This raises issues analogous to those of local average treatment effects (LATE) in the instrumental variables context, in which our main estimates are a weighted average of the underlying treatment effects. Typically these weights might not correspond to our common-sense intuitions or to our desired weighting.[15] This literature typically considers the case where there is (at most) a single event per unit. Sun and Abraham (2021) point out that the overall effect will be a weighted average of the heterogenous effects for different unit types. They show that an auxiliary regression can calculate the implied weights and also propose an alternative estimation method that works to recover a target average treatment effect. Using a different strategy, de Chaisemartin and D'Haultfœuille (2022) propose relying on using not-yet-treated units and the parallel trends assumption to recover estimates of the treatment effects for each treated unit type, which can then be averaged together.

When "Time" Is Not Calendar Time

What if the time variable is not calendar time? This situation can arise in cohort studies: for example Duflo (2001) studies the life-course impact of childhood exposure to school availability in Indonesia based on district and year of birth, and Bailey, Sun, and Timpe (2021) examine the long run impacts of childhood exposure to Head Start in the United States, with treatment based on county and year of birth. These are standard event study analyses, only the time variable is "year of birth" instead of calendar time.

The challenge here is how best to deal with cohort, age, and time (of survey) effects. The basic event study specification requires fixed effects for cohort. Often the outcomes of interest—such as labor market or demographic outcomes—depend

[15] This theme is addressed for ordinary least squares in Angrist (1998) and Sloczynski (2022); for one way fixed effects models in Gibbons, Suárez Serrato, and Urbancic (2019) and Miller, Shenhav, and Grosz (2021); and for difference-in-difference in Goodman-Bacon (2021a), Callaway and Sant'Anna (2021), de Chaisemartin and D'Haultfœuille (2020), and Borusyak, Jaravel, and Spiess (2022).

importantly on age in nonlinear ways. There can also be important calendar-time effects (for example, if some data is collected in a recession). This raises the challenge of age-cohort-time multicollinearity (as discussed in Deaton 2018, pp. 123–127).

There is no avoiding the fact that the analysis becomes complicated here. If theory suggests all three factors— age, cohort, and time—may be important, I recommend including all three sets of dummies.[16] One leading alternative is instead to include a set of two-way interaction dummies: either *age–by–cohort* fixed effects, *cohort–by–time* fixed effects, or *age–by–time* fixed effects. Any one set of these two-way-interactive fixed effects controls for more, but also "uses up" more variation in the data, which raises its own issues. But it seems like good practice to at least include a specification with these two-way-interactive fixed effects as a robustness check.

A final issue is that in producing an overall estimate, we might want to make sure that each cohort is weighted proportional to its population, because the thought experiment of the model centers on the cohorts. However, our data might not naturally reflect those weights, perhaps especially when we are combining data from different-sized datasets.[17] This can set up a choice between improved statistical power (weighting based on the data in our sample) and improved representativeness (weighting based on size of cohorts), and I do not think there is currently a settled "best practice" for these issues. But we should think carefully about how to weight our observations, and not simply take the weights as they are given by the datasets we are using.

Conclusion

Event study models are great! But behind that attractive interpretive graph, researchers are necessarily making decisions. This raises risks of bias due to systematic (if perhaps unconscious) model selection processes, committed by either the researcher or the journal review process. Despite these risks, these decisions are unavoidable. There is no "button to push" that can automate the necessary judgment calls. For now, best practice should be to increase transparency through bringing clarity about the specification decisions made (and the reasons for those decisions) and to discuss robustness to alternative decisions, along with providing both estimation code and (whenever possible) data for replication.

[16] Sometimes researchers include only two sets of these three possible sets of dummy variables, such as dummies for age and for cohort. Presumably this is motivated by the collinear relationship: age = cohort + time. However, among the many coefficients for the three sets of dummy variables, there is only one degree of collinearity. So omitting a full block of dummies—say, leaving out the time dummies—imposes many more restrictions than are necessary.

[17] For example, the outcome data in Bailey, Sun, and Timpe (2021) pool observations from the long-form 2000 Census and the 2001–2018 ACS. The relatively large number of observations in the census means that birth cohorts from 1950 to 1965 are weighted more heavily than later cohorts, relative to their population size.

■ *I thank Gaetano Basso, Colin Cameron, Hilary Hoynes, Giulia Olivero, Marianne Page, Zhuan Pei, Jessamyn Schaller, my students in my PhD Applied Econometrics classes, and the editors, for many helpful comments and conversations. Special thanks to Liz Cascio, who first explained event study models to me in such a way that I felt like I started to "get it."*

References

Alsan, Marcella, and Claudia Goldin. 2019. "Watersheds in Child Mortality: The Role of Effective Water and Sewerage Infrastructure, 1880–1920." *Journal of Political Economy* 127 (2): 586–638.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66 (2): 249–88.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." The Review of Economics and Statistics 60 (1): 47–57.

Bailey, Martha J., Hilary W. Hoynes, Maya Rossin-Slater, and Reed Walker. 2020. "Is the Social Safety Net a Long-Term Investment? Large-Scale Evidence from the Food Stamps Program." NBER Working Paper 26942.

Bailey, Martha J., Shuqiao Sun, and Brenden Timpe. 2021. "Prep School for Poor Kids: The Long-Run Impacts of Head Start on Human Capital and Economic Self-Sufficiency." *American Economic Review* 111 (12): 3963–4001.

Basso, Gaetano, Douglas L. Miller, and Jessamyn Schaller. 2022. "Dynamic Treatment Effects for Empirical Microeconomists: Local Projections and Quasi-experimental Research Designs." Unpublished.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2022. "Revisiting Event Study Designs: Robust and Efficient Estimation." Unpublished.

Bostwick, Valerie, Stefanie Fischer, and Matthew Lang. 2022. "Semesters or Quarters? The Effect of the Academic Calendar on Postsecondary Student Outcomes." *American Economic Journal: Economic Policy* 14 (1): 40–80.

Braun, Matias, and Claudio Raddatz. 2008. "The Politics of Financial Development: Evidence from Trade Liberalization." *Journal of Finance* 63 (3): 1469–1508.

Bronnenberg, Bart J., Jean-Pierre H. Dubé, and Matthew Gentzkow. 2012. "The Evolution of Brand Preferences: Evidence from Consumer Migration." *American Economic Review* 102 (6): 2472–2508.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–230.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–72.

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

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. 2017. "Asymptotic Behavior of a t-Test Robust to Cluster Heterogeneity." *Review of Economics and Statistics* 99 (4): 698–709.

Cellini, Stephanie R., Fernando Ferreira, and Jesse Rothstein. 2010. "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design." *Quarterly Journal of Economics* 125 (1): 215–61.

Chetty, Raj, John N. Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen. 2014. "Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark." *Quarterly Journal of Economics* 129 (3): 1141–1219.

Clarke, Damian, and Kathya Tapia-Schythe. 2021. "Implementing the Panel Event Study." *Stata Journal* 21 (4): 853–84.

Conley, T. G. 1999. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92 (1):

1–45.

Currie, Janet, Henrik Kleven, and Esmée Zwiers. 2020. "Technology and Big Data Are Changing Economics: Mining Text to Track Methods." *AEA Papers and Proceedings* 110: 42–48.

Deaton, Angus. 2018. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Reissue Edition with a New Preface.* Washington, DC: World Bank.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2022. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." NBER Working Paper 29873.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo. 2018. "The Economic Consequences of Hospital Admissions." *American Economic Review* 108 (2): 308– 52.

Dolley, J. C. 1933a. "Characteristics and Procedure of Common-Stock Split-Ups." *Harvard Business Review* 11: 316–26.

Dolley, J. C. 1933b. "Common Stock Split-Ups: Motives and Effects." *Harvard Business Review* 12: 70–81.

Driscoll, John C., and Aart C. Kraay. 1998. "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data." *Review of Economics and Statistics* 80 (4): 549–60.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813.

Finkelstein, Amy, Matthew Gentzkow, Dean Li, and Heidi L. Williams. 2022. "What Drives Risky Prescription Opioid Use? Evidence from Migration. NBER Working Paper 30471.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2016. "Sources of Geographic Variation in Health Care: Evidence from Patient Migration." *Quarterly Journal of Economics* 131 (4): 1681–1726.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. 2019. "Pre-event Trends in the Panel Event-Study Design." *American Economic Review* 109 (9): 3307–38.

Gibbons, Charles E., Juan Carlos Suárez Serrato, and Michael B. Urbancic. 2019. "Broken or Fixed Effects?" *Journal of Econometric Methods* 8 (1): 20170002.

Goodman-Bacon, Andrew. 2018. "Public Insurance and Mortality: Evidence from Medicaid Implementation." *Journal of Political Economy* 126 (1): 216–62.

Goodman-Bacon, Andrew. 2021a. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–77.

Goodman-Bacon, Andrew. 2021b. "The Long-Run Effects of Childhood Insurance Coverage: Medicaid Implementation, Adult Health, and Labor Market Outcomes." *American Economic Review* 111 (8): 2550–93.

Goodman-Bacon, Andrew, and Jamein P. Cunningham. 2019. "Changes in Family Structure and Welfare Participation since the 1960s: The Role of Legal Services." NBER Working Paper 26238.

Imbens, Guido W., and Michal Kolesar. 2016. "Robust Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics* 98 (4): 701–12.

Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan. 1993. "Earnings Losses of Displaced Workers." *American Economic Review* 83 (4): 685–709.

Krolikowski, Pawel. 2018. "Choosing a Control Group for Displaced Workers." *ILR Review* 71 (5): 1232–54.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. "School Finance Reform and the Distribution of Student Achievement." *American Economic Journal: Applied Economics* 10 (2): 1–26.

MacKinlay, A. Craig. 1997. "Event Studies in Economics and Finance." *Journal of Economic Literature* 35 (1): 13–39.

MacKinnon, James G., and Matthew D. Webb. 2017. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32 (2): 233–54.

MacKinnon, James G., and Matthew D. Webb. 2019. "Wild Bootstrap Randomization Inference for Few Treated Clusters." In *The Econometrics of Complex Survey Data: Theory and Applications*, Vol. 39, *Advances in Econometrics*, edited by Kim P. Huynh, David T. Jacho-Chávez, 61–85. Bingley, UK: Emerald Publishing.

Miller, Douglas L. 2023. "Replication data for: An Introductory Guide to Event Study Models." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E185943V1.

Miller, Douglas L., Na'ama Shenhav, and Michel Grosz. 2021. "Selection into Identification, with Application to Head Start." *Journal of Human Resources.* doi: 10.3368/jhr.58.5.0520-10930R1.

Molitor, David. 2018. “The Evolution of Physician Practice Styles: Evidence from Cardiologist Migration.” *American Economic Journal: Economic Policy* 10 (1): 326–56.

Rambachan, Ashesh, and Jonathan Roth. 2023. “A More Credible Approach to Parallel Trends.” *Review of Economic Studies*. https://doi.org/10.1093/restud/rdad018.

Roth, Jonathan. 2022. “Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends.” *American Economic Review: Insights* 4 (3): 305–22.

Sandler, Danielle H., and Ryan Sandler. 2014. “Multiple Event Studies in Public Finance and Labor Economics: A Simulation Study with Applications.” *Journal of Economic and Social Measurement* 39 (1–2): 31–57.

Schmidheiny, Kurt, and Sebastian Siegloch. 2023. “On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization.” *Journal of Applied Econometrics*. doi: 10.1002/jae.2971.

Sloczynski, Tymon. 2022. “Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights.” *Review of Economics and Statistics* 104 (3): 501–09.

Sun, Liyang, and Sarah Abraham. 2021. “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects.” *Journal of Econometrics* 225 (2): 175–99.

Wolfers, Justin. 2006. “Did Unilateral Divorce Laws Raise Divorce Rates? A Reconciliation and New Results.” *American Economic Review* 96 (5): 1802–20.

Young, Alwyn. 2019. “Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results.” *Quarterly Journal of Economics* 134 (2): 557–98.

Journal of Economic Perspectives—Volume 37, Number 2—Spring 2023—Pages 231–246

Retrospectives
Edgar Sydenstricker: Household Equivalence Scales and the Causes of Pellagra

Philip Clarke and Guido Erreygers

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please contact either Beatrice Cherrier, CNRS & CREST, ENSAE-Ecole Polytechnique (beatrice.cherrier@gmail.com) or Joseph Persky, University of Illinois at Chicago (jpersky@uic.edu).

Introduction

The COVID-19 pandemic is not the first time that economists joined forces with epidemiologists to help tackle a major outbreak of disease. Indeed, one of the most successful collaborations just over a century ago involved the causes of the disease of pellagra. While pellagra is almost unknown today, it was endemic in some regions of Europe at least since the middle of the eighteenth century. By the early part of the twentieth century, it was spreading fast in the US South, especially in cotton-growing areas, with about 250,000 cases and 7,000 deaths annually (Bollet 1992). Its main symptoms were sometimes summarized as the three D's: dermatitis, diarrhoea, and dementia. Specifically, pellagra caused severe irritations of the skin (the name comes from the Italian for "sour skin") as well as digestive problems and mental disorders. Pellagra usually peaked in the spring and mainly affected poor

■ *Philip Clarke is Professor of Health Economics, Oxford University, Oxford, United Kingdom. Guido Erreygers is Professor of Economics, University of Antwerp, Antwerp, Belgium. Their email addresses are philip.clarke@ndph.ox.ac.uk and guido.erreygers@uantwerpen.be.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.231.

people, but in the early twentieth century, its cause was unknown. Some claimed it was an infection, similar to diseases like tuberculosis and hookworm. Others argued it was caused by a dietary deficiency, like scurvy and beriberi.

Only in 1937 was the cause of pellagra identified specifically as a dietary deficiency of vitamin B-3, also called nicotinic acid or niacin (for a contemporary review of this discovery, see Elvehjem 1940). Twenty years earlier, a body of research by epidemiologists and economists, based on large-scale fieldwork and the collection of a substantial amount of economic data, established that pellagra was due to a dietary deficiency—although they were not able at that time to pinpoint exactly what. A central figure in this earlier research project was the economist Edgar Sydenstricker. One of the major analytical challenges in studying the economic and dietary causes of pellagra was how to compare across households with similar income levels, but different compositions of adults and children. While economists and other social scientists nowadays routinely apply equivalence factors to adjust household incomes when measuring and analyzing poverty, inequality, and other topics, this was not the case a century ago.[1] Although Sydenstricker did not originate the idea of household equivalence scales to make such comparisons possible, his research offered insights into the basis for such scales, how they might be designed, and their practical importance in contributing to the solution of a major public health issue.

Sydenstricker (1881–1936) had been born in China to missionary parents: his younger sister grew up to be the writer Pearl S. Buck. He moved to the United States in 1896 (for biographical information, see King 1936). He earned a Master of Arts degree at Washington and Lee University in 1902 and studied political economy at the University of Chicago from 1907 to 1908. He then worked for the US Commission on Industrial Relations, where he investigated the labor conditions in American industries. In 1915, he moved to the US Public Health Service, where he was hired as a statistician and did research on the health and socioeconomic status of workers in the New York garment industry as well as on health insurance systems.[2] In 1923 he spent a year at the League of Nations, and in 1928 he became Director of Research at the Milbank Memorial Fund. Throughout his career, Sydenstricker reported frequently about the results of his research on health issues. While working for the US Public Health Service, he published many of his papers in its official journal *Public Health Reports.* Lengthier studies, such as one on health insurance (Warren and Sydenstricker 1916) were published in the *Public Health Bulletin*

[1] Well-known examples of household equivalence measures include the OECD equivalence scale, the OECD-modified scale, and the square root scale (for example, https://www.oecd.org/els/soc/OECD-Note-EquivalenceScales.pdf).

[2] Much of Sydenstricker's work on labor conditions (which is mentioned in the Final Report and Testimony submitted to Congress by the Commission on Industrial Relations, 1916: Vol. I, 14, 22 and 163) was published independently by Lauck and Sydenstricker (1917). It also gave rise to an article in the *Journal of Political Economy* (Sydenstricker 1916).

series.[3] It was not long after his appointment in the US Public Health Service that Sydenstricker became involved in the research about pellagra.

Early Controversy over the Causes of Pellagra

In the early part of the twentieth century, pellagra was widely recognized as an urgent public health issue. In 1909, the US Public Health Service appointed a Pellagra Commission, the first National Conference on Pellagra was held, and the National Association for the Study of Pellagra was founded. In 1912, two wealthy businessmen—Robert M. Thompson and J. H. McFadden—donated a large sum of money for research on pellagra. This allowed what came to be called the Thompson-McFadden Commission to initiate a large-scale epidemiological study in six cotton mill villages in Spartanburg County, South Carolina. After two years of research, the commission came to the conclusion that there was very little evidence for a dietary deficiency theory. It argued that an analysis of location and distance patterns strongly suggested that pellagra was an infectious disease (Siler, Garrison, and McNeal 1914).

In contrast, the epidemiologist Joseph Goldberger (1874–1929), who in 1914 was put in charge of an investigation into pellagra by the US Surgeon General, was convinced that pellagra was caused by a dietary deficiency (Goldberger 1914a). To gather evidence, he set up experiments in an asylum in Georgia and in orphanages in Mississippi, which showed that dietary changes could produce a spectacular drop in pellagra prevalence (Goldberger 1914b; Goldberger, Waring, and Willets 1914; 1915). He also tried to discredit the view that pellagra was an infectious disease by attempting repeatedly, but unsuccessfully, to transmit the disease from pellagra victims to others, including himself and his wife (Goldberger 1916).

This evidence did not suffice to win the debate; the view that pellagra was an infectious disease persisted (Siler, Garrison, and MacNeal 1917a; 1917b). Goldberger therefore decided to launch an epidemiological study quite similar to the one undertaken by the Thompson-McFadden Commission—in effect, redoing the earlier study. The two surveys were not exactly the same (as explained in detail by Mooney, Knox, and Morabia 2014). For example, the set of cotton mill villages was altered: two of the six villages of the original survey were replaced by three other ones. In addition, much more care was given to the collection of data on food supplies, prices, and incomes at the village, household, and individual levels. For this study, Goldberger collaborated closely with George A. Wheeler (1885–1981), who was a doctor and a member of the US Public Health Service, and with Sydenstricker.

[3] Sydenstricker also published in economics and statistics journals (Sydenstricker and King 1921a; 1921b). After he joined the Milbank Memorial Fund, his book *Health and Environment* (Sydenstricker 1933) appeared in the Series of Monographs Prepared under the Direction of the President's Research Committee on Social Trends, initiated by Herbert Hoover and chaired by Wesley Clair Mitchell. In 1974, a selection of Sydenstrickers's public health papers was published under the title *The Challenge of Facts* (Sydenstricker 1974).

They published their findings in a series of papers (Goldberger, Wheeler, and Sydenstricker 1918; 1920a; 1920b; 1920c).

The Need for a Household Equivalence Scale

From a very early stage—even before the start of the Goldberger survey—Sydenstricker (1915, p. 3132) had drawn attention to the economic aspects of the issue when he alluded to "the possible relation of the incidence of pellagra to certain conditions of an economic character." He had first-hand knowledge of labor and industrial conditions. He was aware of the evidence that, in comparison to richer families, poor families spent a greater proportion of their family income on food, and that the per capita consumption of milk, eggs, and meat tended to rise with income. Moreover, he pointed out that the typical wage earner in the South had less purchasing power than the one in the North, and that in certain communities of the South the food supply was often limited.

Sydenstricker supervised the collection of the income data in the context of Goldberger's 1916 survey on pellagra in the textile mill communities of South Carolina. Two issues were crucially important: obtaining reliable estimates of total household incomes and making an appropriate adjustment for household composition. They used the cotton mill payroll records to supplement and verify the data on self-reported income (Goldberger, Wheeler and Sydenstricker 1920c, pp. 2680–1). It soon became obvious, however, that socioeconomic status was much more varied than household income, as the size and composition of households varied markedly. The authors wrote (pp. 2682–3),

> The average total annual cash income of all of the families for which income data were secured was about $700, and relatively few had annual incomes of over $1,000. Thus the range of total income was relatively small and the families were, from this point of view, fairly homogeneous. They differed, however, very markedly in size and with respect to the age and sex of their members. Manifestly it was improper to classify, for example, a family whose half-month's income was $40, and was composed of only a man and his wife, with one whose half-month's income was also $40, but was composed of a man, his wife, and several dependent children.

The question was how to compare families with different age-sex compositions. Previous studies examining the epidemiology of pellagra such as Jobling and Petersen (1917) had used per capita income as a measure of socioeconomic status. Although Goldberger, Wheeler, and Sydenstricker (1920c, p. 2683) acknowledged that dividing household income by the number of persons was better than making no adjustments at all, they argued that the use of per capita income remained inaccurate because it entirely neglected differences in age and sex: "It appeared advisable, therefore, to employ a common denominator to which the individuals of both sexes and of all ages could be reduced in order to obtain a more accurately representative method of expressing the relative size of the families to be compared."

This search for a common denominator was the starting point of Sydenstricker's work on the development and use of equivalence scales, which would enable him to assess how the incidence of diseases like pellagra varied by socioeconomic status. For the derivation of equivalence scales, Sydenstricker collaborated with Willford I. King, who joined Sydenstricker at the US Public Health Service in 1917 before taking up a position at the National Bureau of Economic Research in 1920. King had been instructor and assistant professor of political economy at the University of Wisconsin, where he also earned his PhD (as reported in *New York Times* 1962).[4]

The three articles Sydenstricker and King published in 1920 and 1921 spell out how their research is connected to the pellagra studies and what they wanted to achieve. Their main motivation was to study the association between poverty and ill-health in a quantitative way (Sydenstricker and King 1920, p. 2830). Family income as such was too crude a measure: it ignored differences in size and differences in composition (sex and age) of different households. Per capita income was an improvement, but nevertheless they regarded it as being "subject to a high degree of error" (Sydenstricker and King 1921a, p. 583) and so saw the need to develop a more refined measure of socioeconomic status. The alternative they proposed was to express the size and composition of different families into comparable units and then to divide each family's income by the corresponding number of units. Sydenstricker and King acknowledged that this procedure was not new; for earlier examples, they referred to the work of the Englishman David Davies at the end of the eighteenth century, the German Ernst Engel in the nineteenth century (who suggested the name "quet" for the units, after the Belgian statistician Adolphe Quetelet), and the Englishman Seebohm Rowntree. In the United States, a household equivalence scale had been developed by Wilbur Olin Atwater, the American chemist who allegedly introduced the word "calorie" into US English (Hargrove 2006). Through experimentation, Atwater examined the amount of energy contained in different types of foods and calculated estimates of the number of calories that needed to be consumed by persons of different ages and gender to sustain life (Carpenter 1994). With the Atwater scale, the size of the family could be expressed in Adult Male Units, using males over the age of 16 that have the highest calorie needs as the reference point. It was by dividing household income by the number of Adult Male Units that Sydenstricker and his associates classified families according to income in

[4] King (1915) had previously published a study on the distribution of income and wealth in the United States. While working with Sydenstricker, he simultaneously wrote a review of a study on disabling sickness by Sydenstricker, Wheeler, and Goldberger (1918) in the *American Economic Review* (King 1919). This article clearly shows that King shared many of Sydenstricker's views on the relationships between income and health and on the necessity to adjust income for household composition.

their analysis of the incidence of disabling sickness (Sydenstricker, Wheeler, and Goldberger 1918).

However, Sydenstricker and King (1920, pp. 2834–5) were not completely satisfied with the Atwater scale. They argued that the number of units had to also take into account the needs for nonfood items, and that these might be different from the needs for food. They wrote:

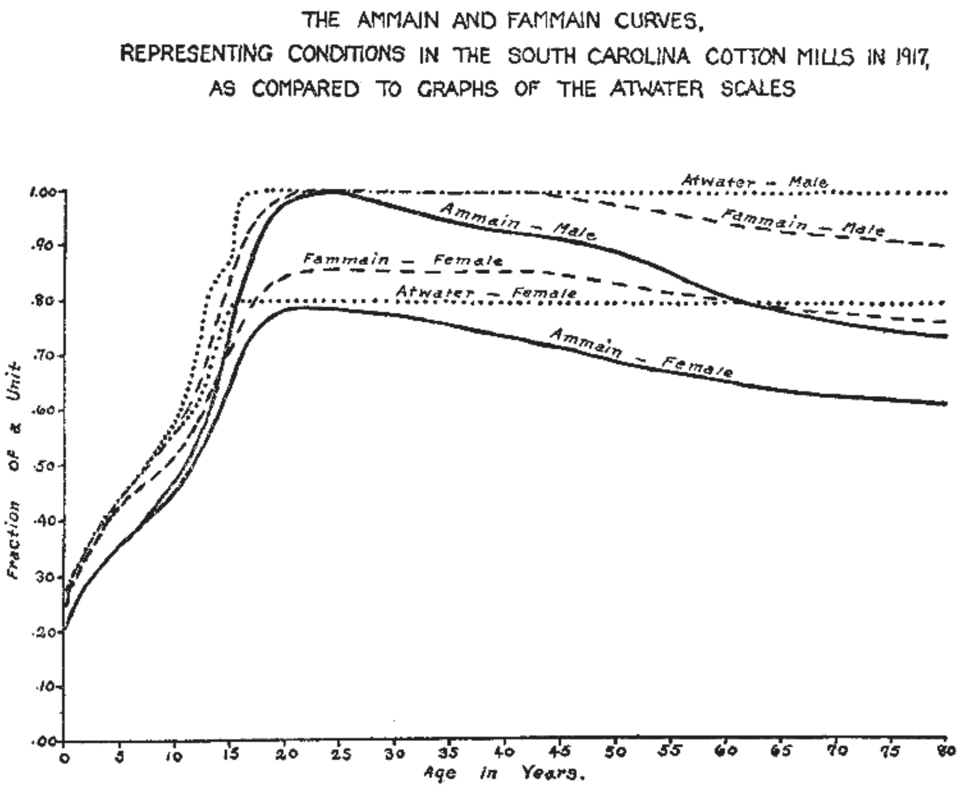
> [T]he assumption in the use of this scale was that the expenditures for individuals varied according to sex and age in the same proportion as their food requirements. The assumption was by no means as accurate as could be desired . . . Accordingly, it was determined to use as the basis for classifying families a unit which would express as far as possible the relative differences in all economic wants among persons of different sexes and ages.

In 1917, they had a new opportunity to collect data in the context of the ongoing pellagra surveys. The design of the new survey was modified to remedy some of the weaknesses of the 1916 survey; for example, no information had been collected in the 1916 version of the survey on home-produced food. The household expenditure data were then analyzed to detect differences in expenditure patterns according to sex and age, with the aim of measuring more accurately the relative well-being of different families. As described in Sydenstricker and King 1920, p. 2835),

> In analyzing the data the problem was found to consist of two parts: First, the derivation of correct curves showing food expenditures, and, second, the obtaining of similar curves for other expenditures. The reason that this division of the problem was necessary lay in the fact that the food used, including that purchased and that produced at home, was recorded only for the family as a whole, and it was entirely impracticable to secure separate records for individuals. On the other hand, expenditures for such articles as clothing, medical care, tobacco, amusement, etc., might actually be estimated for the different individuals in the family.

Because their analysis was based on observed expenditures rather than on calculations of food requirements, they decided to use a new name for the unit of their adjusted scale. They settled on the *fammain,* short for *food for adult male maintenance*. For the derivation of the fammain, they used the food expenditure records of about 1,500 families. To analyze the nonfood expenditures, they needed complete household expenditure records, and these were not so readily available. On the basis of budget data of about 140 families, they nevertheless derived a scale representative for the differences in total expenditure according to age and sex. The unit of that scale was called the *ammain,* short for *adult male maintenance*. While the main purpose of the fammain equivalence scale was to

Figure 1
Comparison of Different Scales Undertaken by Sydenstricker and King



Source: Sydenstricker and King (1921b, p. 850).
Notes: The Atwater scales for male and female are shown with dotted lines. The fammain scales for male and female are shown with dashed lines. The ammain scales for male and female are shown with solid lines.

enable a comparison of the food supplies of different families, the goal of the ammain equivalence scale was to allow a much broader comparison: it took into account all household expenditures and was meant as an instrument to measure how well-off families actually were.

A comparison of the Atwater, fammain, and ammain scales is reproduced in Figure 1. All of the measures show a similar rising pattern through childhood, culminating in a peak around age 18—although the peak is higher for males than for females. Having the peak for men happen at 1.0 arises out of the construction of the figure. At that point, the Atwater scale remains constant for the rest of the lifespan. However, Sydenstricker and King found that the values of the fammain and ammain scales tended to decline for older adults, with the ammain values (that is, all expenditures, not just food) falling faster. Sydenstricker and King were convinced that these two scales, derived from observed household expenditure data, were more accurate than the Atwater scale, which was based on estimated calorie requirements. On the whole, however, the numerical values of the different scales are relatively similar.

Using Equivalent Income to Understand the Cause of Pellagra

Although Sydenstricker and King used the data from the pellagra surveys to calculate their fammain and ammain scales, the scales were not applied in the pellagra studies of Goldberger, Wheeler, and Sydenstricker (1918; 1920a; 1920b; 1920c). Sydenstricker probably thought that the methodology needed further testing and refinement before it could be put into practice. In the conclusion of their methodological paper in the *Quarterly Publications of the American Statistical Association*, Sydenstricker and King (1921b, p. 856) wrote that they presented their study "with a full realization that it is only a beginning and not a finality in this field." Sydenstricker's cautious stance explains why, in the pellagra studies done by the US Public Health Service, equivalent incomes were calculated using the Atwater scale. Goldberger, Wheeler, and Sydenstricker (1920c, p. 2683) said they did so in "the absence of a better common denominator for this purpose . . . The assumption is by no means as accurate as could be desired," they wrote, but "a scale based on food requirements alone is obviously very much more accurate than one omitting any consideration whatsoever of the number, sex, and age of the individuals composing the families."

Even when equivalent incomes were estimated with this imperfect device, the analysis of the data revealed that pellagra and equivalent income were highly correlated. The table shown in Figure 2 indicates that both the crude incidence rate of pellagra (based on raw data) and the rate adjusted for age and sex differences are characterised by a strong association with the level of equivalent income. This is one of the rare occasions when Sydenstricker and his coauthors (p. 2686) quantified the degree of correlation: "Upon the basis of the average half-month income per adult male unit for each of the income classes and the corresponding pellagra rate per 1,000 persons, the Pearsonian coefficient of correlation is –0.91 ± 0.05."

Moreover, comparisons between households with and without pellagra revealed striking differences in dietary habits. For a list of food items, Goldberger, Wheeler, and Sydenstricker (1920a, pp. 669–72) calculated the available supply in grams per adult male unit. To highlight the differences, they set the quantity of the average diet for nonpellagra households of the highest income class equal to an index number of 100, as shown in Figure 3. They then compared the diet of this group of households to those of three other groups of households: nonpellagra households of the lowest income class, pellagra households of the lowest income class, and households with at least two cases of pellagra. These calculations indicated which food items might be associated with the incidence of pellagra. In the pellagra households with low-income levels, the quantities consumed of goods to the left-hand side of the

Figure 2
Pellagra Incidence per Income Class

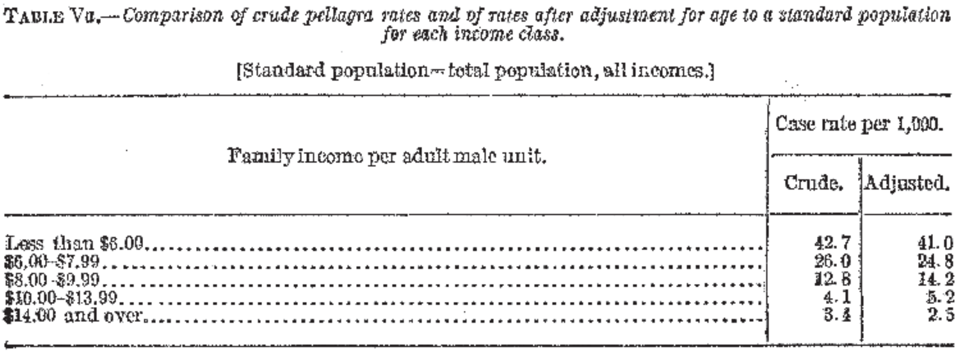


Table VII.—*Comparison of crude pellagra rates and of rates after adjustment for age to a standard population for each income class.*

[Standard population—total population, all incomes.]

Family income per adult male unit.	Case rate per 1,000.	
	Crude.	Adjusted.
Less than $6.00	42.7	41.0
$6.00–$7.99	26.0	24.8
$8.00–$9.99	12.8	14.2
$10.00–$13.99	4.1	5.2
$14.00 and over	3.4	2.5

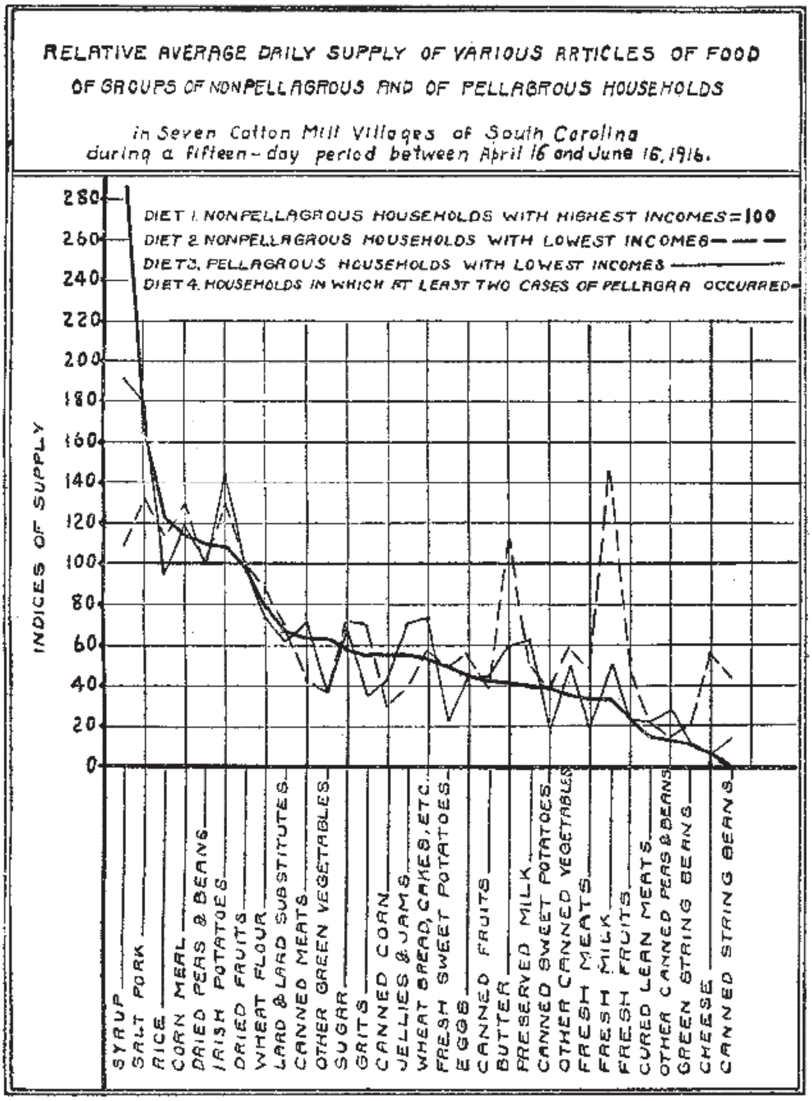
Source: Goldberger, Wheeler, and Sydenstricker (1920c, p. 2687).
Note: Since the authors found that the incidence of pellagra varied according to age, they recalculated the incidence rates for each income class assuming that every class had the same age distribution. A comparison of the crude and adjusted incidence rates shows that the steep income gradient is not a result of the difference in age distribution between the income classes.

figure, like syrup, salt pork, rice, and corn meal, were actually higher than in the nonpellagra high-income households. However, in nonpellagra households with low-income levels, the consumption of "the group comprising the animal protein foods, namely, the lean meats, milk, butter, cheese, and eggs" was clearly higher than in pellagra households. These foods are all major sources of dietary niacin—vitamin B-3. The authors did not know that specific detail, but they nonetheless wrote (p. 681) that "the freedom from pellagra of the nonpellagrous households can be considered as significantly associated with a more liberal supply from but this one group of foods."

Based on this data, Goldberger, Wheeler, and Sydenstricker (1920a, p. 702) argued that their findings constituted "additional evidence of the controlling influence of diet in the prevention and the causation of the disease." They arrived at the conclusion that while pellagra was negatively associated with socioeconomic status, this was due primarily to differences in the diet of low- and high-income households. The few cases occurring in higher income households were likely explainable by them having a diet that was similar to the diet of those at risk of pellagra in low-income households. The clear connection between equivalent income and pellagra incidence also meant that economic conditions such as areas of poverty or cycles of booms and busts, which affect incomes via changes in prices, wages and employment levels, were bound to influence the incidence rates of the disease (Goldberger, Wheeler, and Sydenstricker 1920c, pp. 2709–10).

Although the US Public Health Service applied the Atwater scale rather than the fammain and ammain scales in the analysis of the pellagra survey data, the use of *any* equivalence scale, albeit an imperfect one, was crucially important to

Figure 3
Diet Comparisons



Source: Goldberger, Wheeler, and Sydenstricker (1920a, p. 677).
Notes: For nonpellagrous households with the highest incomes, the quantity per person of each food listed on the horizontal axis is indexed to equal 100. For nonpellagrous families with lowest incomes, the quantity of each food is shown by the wide-dashed line. For pellagrous families with lowest incomes, this is shown by a solid line, and for families with at least two cases of pellagra by a bold line.

provide evidence for the dietary deficiency theory and to highlight the relevance of economic factors. It was instrumental to stratify the prevalence of the disease and to study patterns of food consumption. The analysis showed both dramatic variation in the prevalence of the disease—like a 16-fold difference in the rates of pellagra between the lowest and highest equivalized income categories—as well as considerable variation in diet across these income categories. Such evidence was central to the US Public Health Service making the case that pellagra, rather than being an infectious disease, was linked to some type of dietary issue (later identified as low levels of niacin) and strongly associated with poverty.

Two caveats to these findings are worth noting. First, while some epidemiologists have sought to reanalyse both the work of the Thompson-McFadden Commission and that of the US Public Health Service, such efforts are limited by not having access to original individual level data (Mooney, Knox, and Morabia 2014). Hence, it is difficult to now isolate the contribution of using equalized income over other measures of socioeconomic status in helping illuminate the causes of pellagra at this time. Although it is worth noting that previous epidemiological studies such as Jobling and Petersen (1917) used per capita income, they did not find the same level of correlation between income and cases of pellagra and in fact reached opposite conclusions regarding its cause (that is, incorrectly finding that pellagra was an infectious disease).

Second, the pellagra fieldwork undertaken by the US Public Health Service involved by design a largely homogenous sample and this involved excluding some occupations (like mill managers) as well as African American families. Marks (2003) has argued that Sydenstricker, in focusing his effort on understanding patterns of disease by income, ignored important racial and gender health disparities, which have been examined by a later generation of economists (for example, Myrdal 1944; Muller 1990).

What Can We Learn from Sydenstricker's Contributions Today?

Sydenstricker and King (1921a, p. 593) emphasized the potential uses of their equivalence scale:

> The advantages of reducing family income to a per ammain basis are obvious and striking. By this process all families are made readily comparable as to income, entirely regardless of their respective age and sex compositions or of the number of persons composing them. This mode of procedure substitutes precise for haphazard methods, brings order out of chaos, and, in fact, offers the first plan which really permits of the income classification of all of the families in any locality with any reasonable degree of precision.

They suggested that "the Bureau of Labor Statistics, the Treasury Department, or some other interested branch of the government derive official ammain scales based upon an average representing all classes of the population."

In practice, however, there was only modest use of the newly developed ammain scale. Wiehl and Sydenstricker (1924) applied it in the analysis of the incidence of disabling sickness in a range of cotton mill towns of South Carolina, using data from the second wave of the pellagra survey of 1917. This was a follow-up study on a previous analysis by Sydenstricker, Wheeler, and Goldberger (1918), which used data from the more restricted first wave of the pellagra survey of 1916 and applied the Atwater equivalence scale. It is not possible to compare the two studies directly, because they report patterns of disease across different income groups, but the results from both studies show a strong gradient in disease prevalence across income groups within these communities. In addition, Sydenstricker, King, and Wiehl (1924) employed the ammain scale in a study of life cycle variations in the income of wage earners. King (1925) referred to the scale in a paper on the measurement of income and wealth.

In Sydenstricker's time, the US Public Health Service invested in additional fieldwork to collect economic data that could be used to construct more refined equivalence scales. While Sydenstricker and the Public Health Service in subsequent years continued to focus on the relationship between income and health, simpler measures such as per capita income were often regarded as sufficient (Perrott, Collins, and Sydenstricker 1933, p. 1256). As time went on, even the idea of using data on income for the study of disease fell into disuse by public health researchers. As one prominent example, the massive Framingham Heart study was set up by the US Public Health Service in the late 1940s with dedicated funding from Congress to look at the potential causes of heart disease (Mahmood et al. 2014). However, household income was not collected as part of the original Framingham study (Oppenheimer 2005, p. 608). Only information on educational attainment and place of residence was gathered. Apparently, the concern was that asking participants about income might offend them and hence compromise the collection of predominantly clinical data, which was seen as the main purpose of the study (Dawber et al. 1959). Not being able to stratify cardiovascular risk by socioeconomic factors such as income has had a legacy, as the risk-prediction tools developed from Framingham data have been shown to under-predict cardiovascular risk in low-income and less-educated populations (Fiscella, Tancredi, and Franks 2009).

According to Krieger and Fee (1996), the reorganisation of the US Public Health Service and other federal agencies as well as changing attitudes during the Cold War explain why the focus of public health research shifted away from quantifying socioeconomic status. Less attention was paid to the measurement of income gradients and to the use of equivalence scales. For example, the second and third wave of the National Health Examination Survey (Roberts 1973) makes no adjustment for household size or composition when looking at the gradient across income. Many recent summaries of the National Health and Nutrition Examination Survey (NHANES), the most recent incarnation of the health examination surveys, do not report disparities by income at all (for example, Aguilar et al. 2015). A number of recent epidemiological studies do not often use equivalent income and instead rely on measures such as education and occupational class (for example, D'Errico et al. 2017).

Meanwhile, household equivalence scales are often used in the economics literature, although their shape has changed from the research of a century ago.

Among economists and other social scientists, the ammain scale did find its way into other applications, such as Alvin Hansen's (1922) analysis of inflation during the First World War. A search of Google Scholar indicates that Sydenstricker and King's *Journal of Political Economy* article (1921a) has just four citations. Their article in the *Quarterly Publications of the American Statistical Association* (Sydenstricker and King 1921b) received more attention: 118 citations due to its extensive analysis by the sociologist William F. Ogburn (1931) and a discussion by Milton Friedman (1952). Although Friedman (p. 13) very much appreciated "their excellent article," he noted that their approach has two difficulties: it is not always possible to allocate expenditure to individual family members, and some expenditures (like rent) are in the form of "overhead costs," which are fixed and thus will vary by the number in the family. Friedman (p. 14) also raised what he saw as a fundamental problem in obtaining such a scale: in order to determine if families are "equally well off," you need to have a measure of each individual's expenditure requirements. This he regarded as a "vicious circle: we cannot get a satisfactory ammain scale unless we have one to begin with."

A modest rehabilitation of the methodological research of Sydenstricker and King can be found in the work of the economist Sigbert Prais. Although Prais (1953, p. 792) developed his methods for estimating equivalence scales from family budgets without previously knowing their work, he noted that his approach had "many formal analogies with the method of Sydenstricker and King given in a fundamental paper written in 1921 which dealt with many of the problems of principle and practice encountered in this work, but which was virtually ignored in the subsequent literature." Therefore, the work of Sydenstricker and King can be regarded as anticipating the Prais-Houthakker model for measuring equivalence scales (Prais and Houthakker 1955; Muellbauer 1980; Barten 1987).

As Lewbel and Pendakur (2018) have noted, equivalence scales have found a wide range of applications in economic research: assessing the extent of poverty and inequality, calibrating the level of social benefits, measuring the costs of raising children, and so forth. By contrast, they have rarely been used in modern epidemiological studies. We feel that it is time to learn some lessons from history and to adopt Sydenstricker's broader approach in studies attempting to understand public health issues. More generally, involving economists in the data collection and analysis of large epidemiological studies could provide greater insights into the causes and potential interventions to address a wide range of diseases.

■ *This paper is based on work previously presented at the 43rd Annual Meeting of the History of Economics Society (Duke University 2016) and at the 24th Annual Conference of the European Society for the History of Economic Thought (University of National and World Economy, Sofia, 2021). A paper dealing more generally with Sydenstricker's work has recently been published by the European Journal of the History of Economic Thought (Clarke and Erreygers 2022).*

References

Aguilar, Maria, Taft Bhuket, Sharon Torres, Benny Liu and Robert J. Wong. 2015. “Prevalence of the Metabolic Syndrome in the United States, 2003–2012.” *Journal of the American Medical Association* 313 (19): 1973–4.

Avery, Christopher, William Bossert, Adam Clark, Glenn Ellison, and Sara Fisher Ellison. 2020. “An Economist’s Guide to Epidemiology Models of Infectious Disease.” *Journal of Economic Perspectives* 34 (4): 79–104.

Barten, Anton. P. 1987. “Household Budgets.” In *The New Palgrave: A Dictionary of Economics*, Vol. 2, edited by John Eatwell, Murray Milgate, and Peter Newman, 672–7. London: Palgrave-Macmillan.

Bollet, Alfred J. 1992. “Politics and Pellagra: The Epidemic of Pellagra in the U.S. in the Early Twentieth Century.” *Yale Journal of Biology and Medicine* 65 (3): 211–21.

Carpenter, Kenneth J. 1994. “The Life and Times of W. O. Atwater (1844–1907).” *Journal of Nutrition* 124 (S9): S1707–14.

Clarke, Philip, and Guido Erreygers. 2022. “Edgar Sydenstricker, a Pioneer of Health Economics.” *European Journal of the History of Economic Thought* 29 (6): 1066–88.

Commission on Industrial Relations. 1916. *Industrial Relations: Final Report and Testimony*. Washington, DC: Government Printing Office.

Dawber, Thomas R., William B. Kannel, Nicholas Revotskie, Joseph Stokes III, Abraham Kagan, and Tavia Gordon. 1959. “Some Factors Associated with the Development of Coronary Heart Disease: Six Years’ Follow-Up Experience in the Framingham Study.” *American Journal of Public Health* 49 (10): 1349–56.

D’Errico, Angelo, Fulvio Ricceri, Silvia Stringhini, Cristian Carmeli, Mika Kivimaki, Mel Bartley, Cathal McCrory et al. 2017. “Socioeconomic Indicators in Epidemiologic Research: A Practical Example from the LIFEPATH Study.” *PLoS ONE* 12 (5): 1–32.

Di Guilmi, Corrado, Georgios Galanis, and Georgios Baskozos. 2022. “A Behavioural SIR Model: Implications for Physical Distancing Decisions.” *Review of Behavioral Economics* 9 (1): 45–63.

Elvehjem, Conrad A. 1940. “Relation of Nicotinic Acid to Pellagra.” *Physiological Reviews* 20 (2): 249–71.

Fiscella, Kevin, Daniel Tancredi, and Peter Franks. 2009. “Adding Socioeconomic Status to Framingham Scoring to Reduce Disparities in Coronary Risk Assessment.” *American Heart Journal* 157 (6): 988–94.

Friedman, Milton. 1952. “A Method of Comparing Incomes of Families Differing in Composition.” In *Studies in Income and Wealth*, Vol. 15, 9–24. New York: NBER.

Goenka, Aditya, Lin Liu, and Manh-Hung Nguyen. 2021. “SIR Economic Epidemiological Models with Disease Induced Mortality.” *Journal of Mathematical Economics* 93 (C): 102476.

Goldberger, Joseph. 1914a. “The Etiology of Pellagra: The Significance of Certain Epidemiological Observations with Respect Thereto.” *Public Health Reports* 29 (26): 1683–6.

Goldberger, Joseph. 1914b. “The Cause and Prevention of Pellagra.” *Public Health Reports* 29 (37): 2354–7.

Goldberger, Joseph. 1916. “The Transmissibility of Pellagra: Experimental Attempts at Transmission to the Human Subject.” *Public Health Reports* 31 (46): 3159–73.

Goldberger, Joseph, C. H. Waring, and David G. Willets. 1914. “The Treatment and Prevention of Pellagra.” *Public Health Reports* 29 (43): 2821–5.

Goldberger, Joseph, C. H. Waring, and David G. Willets. 1915. “The Prevention of Pellagra: A Test of Diet among Institutional Inmates.” *Public Health Reports* 30 (43): 3117–31.

Goldberger, Joseph, George A. Wheeler, and Edgar Sydenstricker. 1918. “A Study of the Diet of Non-pellagrous and of Pellagrous Households in Textile Mill Communities in South Carolina in 1916.” *Journal of the American Medical Association* 71 (12): 944–49.

Goldberger, Joseph, George A. Wheeler, and Edgar Sydenstricker. 1920a. “A Study of the Relation of Diet to Pellagra Incidence in Seven Textile-Mill Communities of South Carolina in 1916.” *Public Health Reports* 35 (12): 648–713.

Goldberger, Joseph, George A. Wheeler, and Edgar Sydenstricker. 1920b. “Pellagra Incidence in Relation to Sex, Age, Season, Occupation, and ‘Disabling Sickness’ in Seven Cotton-Mill Villages of South Carolina during 1916.” *Public Health Reports* 35 (28): 1650–64.

Goldberger, Joseph, George A. Wheeler, and Edgar Sydenstricker. 1920c. “A Study of the Relation of Family Income and Other Economic Factors to Pellagra Incidence in Seven Cotton-Mill Villages of South Carolina in 1916.” *Public Health Reports* 35 (46): 2673–714.

Hansen, Alvin H. 1922. “The Buying Power of Labor during the War.” *Journal of the American Statistical*

Association 18 (137): 56–66.

Hargrove, James L. 2006. "History of the Calorie in Nutrition." *Journal of Nutrition* 136 (12): 2957–61.

Jobling, James W. and William F. Petersen. 1917. "The Epidemiology of Pellagra in Nashville, Tennessee, II." *Journal of Infectious Diseases* 21 (2): 109–31.

King, Willford I. 1915. *The Wealth and Income of the People of the United States.* New York: Macmillan.

King, Willford I. 1919. Review of *Disabling Sickness among the Population of Seven Cotton Mill Villages of South Carolina in Relation to Family Income*, by Edgar Sydenstricker, George A. Wheeler, and Joseph Goldberger. *American Economic Review* 9 (2): 375–77.

King, Willford I. 1925. "Income and Wealth." *American Economic Review* 15 (3): 457–74.

King, Willford I. 1936. "Edgar Sydenstricker." *Journal of the American Statistical Association* 31 (194): 411–4.

Krieger, Nancy, and Elizabeth Fee. 1996. "Measuring Social Inequalities in Health in the United States: A Historical Review, 1900–1950." *International Journal of Health Services* 26 (3): 391–418.

Lauck, William Jett, and Edgar Sydenstricker. 1917. *Conditions of Labor in American Industries: A Summarization of the Results of Recent Investigations.* New York and London: Funk & Wagnalls.

Lewbel, Andrew, and Krishna Pendakur. 2018. "Equivalence Scales." In *The New Palgrave Dictionary of Economics*, 3873–8. London: Macmillan.

Mahmood, Syed S., Daniel Levy, Ramachandran S. Vasan, and Thomas J. Wang. 2014. "The Framingham Heart Study and the Epidemiology of Cardiovascular Disease: A Historical Perspective." *The Lancet* 383 (9921): 999–1008.

Marks, Harry M. 2003. "Epidemiologists Explain Pellagra. Gender, Race, and Political Economy in the Work of Edgar Sydenstricker." *Journal of the History of Medicine and Allied Sciences* 58 (1): 34–55.

Mooney, Stephen J., Justin Knox, and Alfredo Morabia. 2014. "The Thompson-McFadden Commission and Joseph Goldberger: Contrasting 2 Historical Investigations of Pellagra in Cotton Mill Villages in South Carolina." *American Journal of Epidemiology* 180 (3): 235–44.

Muellbauer, John. 1980. "The Estimation of the Prais-Houthakker Model of Equivalence Scales." *Econometrica* 48 (1): 153–76.

Muller, Charlotte F. 1990. *Health Care and Gender.* New York: Russell Sage Foundation.

Murray, Eleanor J. 2020. "Epidemiology's Time of Need: COVID-19 Calls for Epidemic-Related Economics." *Journal of Economic Perspectives* 34 (4): 105–20.

Myrdal, Gunnar. 1944. *An American Dilemma: The Negro Problem and Modern Democracy*, New York: Harper & Brothers.

New York Times. 1962. "Dr. Willford King Is Dead at 82; Economist, Ex-N.Y.U. Professor." October 19, p. 20.

Ogburn, William F. 1931. "A Device for Measuring the Size of Families, Invented by Edward Sydenstricker and W.I. King." In *Methods in Social Science: A Case Book*, edited by Stuart A. Rice, 210–19. Chicago: University of Chicago Press.

Oppenheimer, Gerald M. 2005. "Becoming the Framingham Study 1947-1950." *American Journal of Public Health* 95 (4): 602–10.

Perrott, G. S. J, Selwyn D. Collins, and Edgar Sydenstricker. 1933. "Sickness and the Economic Depression: Preliminary Report on Illness in Families of Wage Earners in Birmingham, Detroit, and Pittsburgh." *Public Health Reports* 48 (41): 1251–64.

Philipson, Tomas. 2000. "Economic Epidemiology and Infectious Diseases." In *Handbook of Health Economics*, Vol. 1B, edited by Anthony J. Culyer and Joseph P. Newhouse, 1761–99. Amsterdam: Elsevier.

Prais, Sigbert J. 1953. "The Estimation of Equivalent-Adult Scales from Family Budgets." *Economic Journal* 63 (252): 791–810.

Prais, Sigbert J., and Hendrik S. Houthakker. 1955. *The Analysis of Family Budgets.* Cambridge: Cambridge University Press.

Roberts, Jean. 1973. "Examination and Health History Findings among Children and Youths 6-17 Years." *Vital Health Statistics* 129: 1–76.

Schereschewsky, J. W., Warren B. S., and Sydenstricker Edgar. 1916. "Health of Garment Workers. The Relation of Economic Status to Health." *Public Health Reports* 31 (21): 1298–1305.

Siler, Joseph F., P. E. Garrison and W. J. MacNeal. 1914. "A Statistical Study of the Relation of Pellagra to Use of Certain Foods and to Location of Domicile in Six Selected Industrial Communities." *Archives of Internal Medicine* 14 (3): 293–373.

Siler, Joseph F., P. E. Garrison and W. J. MacNeal. 1917a. "Relation of Pellagra to Location of Domicile in Spartan Mills, S. C., and the Adjacent District." *Archives of Internal Medicine* 20 (2): 198–315.

Siler, Joseph F., P. E. Garrison and W. J. MacNeal. 1917b. "The Relation of Pellagra to Location of Domicile in Inman Mills, Inman, S. C." *Archives of Internal Medicine* 20 (4): 521–74.

Sydenstricker, Edgar. 1915. "The Prevalence of Pellagra: Its Possible Relation to the Rise in the Cost of Food." *Public Health Reports* 30 (43): 3132–48.

Sydenstricker, Edgar. 1916. "The Settlement of Disputes under Agreements in the Anthracite Industry." *Journal of Political Economy* 24 (3): 254–83.

Sydenstricker, Edgar. 1933. *Health and Environment*. New York: McGraw-Hill.

Sydenstricker, Edgar. 1974. *The Challenge of Facts: Selected Public Health Papers of Edgar Sydenstricker*, edited by Richard V. Kasius, New York: Milbank Memorial Fund.

Sydenstricker, Edgar, and Willford I. King. 1920. "A Method of Classifying Families According to Incomes in Studies of Disease Prevalence." *Public Health Reports* 35 (48): 2829–46.

Sydenstricker, Edgar, and Willford I. King. 1921a. "The Classification of the Population According to Income." *Journal of Political Economy* 29 (7): 571–94.

Sydenstricker, Edgar, and Willford I. King. 1921b. "The Measurement of the Relative Economic Status of Families." *Quarterly Publications of the American Statistical Association* 17 (135): 842–57.

Sydenstricker, Edgar, Willford I. King, and Dorothy Wiehl. 1924. "The Income Cycle in the Life of the Wage-Earner." *Public Health Reports* 39 (34): 2133–40.

Sydenstricker, Edgar, George A. Wheeler, and Joseph Goldberger. 1918. "Disabling Sickness among the Population of Seven Cotton Mill Villages of South Carolina in Relation to Family Income." *Public Health Reports* 33 (47): 2038–51.

Truett, Jeanne, Jerome Cornfield, and William Kannel. 1967. "A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham." *Journal of Chronic Diseases* 20 (7): 511–24.

Warren, B. S, and Edgar Sydenstricker. 1916. "Health Insurance. Its Relation to the Public Health." *Public Health Bulletin* (76). Washington: Government Publishing Office.

Wiehl, Dorothy, and Edgar Sydenstricker. 1924. "Disabling Sickness in Cotton Mill Communities of South Carolina in 1917: A Study of Sickness Prevalence and Absenteeism, as Recorded in Repeated Canvasses, in Relation to Seasonal Variation, Duration, Sex, Age, and Family Income." *Public Health Reports* 39 (24): 1417–43.

Journal of Economic Perspectives—Volume 37, Number 2—Spring 2023—Pages 247–254

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by e-mail at taylort@macalester.edu, or c/o Journal of Economic Perspectives, *Macalester College, 1600 Grand Ave., Saint Paul, MN 55105.*

Smorgasbord

Stijn Van Nieuwerburgh delivered the 2023 Presidential Address to the American Real Estate and Urban Economics Association on the topic “The remote work revolution: Impact on real estate values and the urban environment” (*Real Estate Economics*, January 2023, pp. 7–48, https://onlinelibrary.wiley.com/doi/10.1111/1540-6229.12422). “Born out of necessity, remote work now appears to have taken hold as a permanent feature of modern labor markets. It is a benefit that employees enjoy and are willing to pay for. Their tolerance for commuting appears to be permanently reduced. Having experienced the flexibility that comes with working from home (WFH), the genie is out of the bottle. Firm managers too have come around to see its virtues, often in the form of higher productivity

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.37.2.247.

and profits, and have adjusted their own expectations about the number of days they expect employees to be in the office. Several firms have gone fully remote, while most others have moved to a hybrid work schedule of 2–3 days in the office. Various indicators of office demand appear to have stabilized at levels far below their prepandemic high-water marks. . . . In this article, I review what we know so far about these dramatic changes in residential and commercial real estate markets."

Alexander J. Field runs counter to conventional wisdom in "The decline of US manufacturing productivity between 1941 and 1948" (*Economic History Review*, published online January 16, 2023, https://onlinelibrary.wiley.com/doi/full/10.1111/ehr.13239). From the abstract: "The view that war benefits potential output has been influential in treatments of US mobilization for the Second World War, where it has been largely premised on the benefits of learning by doing in producing military durables. If the thesis that war benefits aggregate supply is correct, it is indeed within manufacturing that we should most likely see its effects. Total factor productivity within the sector in fact fell at a rate of −1.4 percent per year between 1941 and 1948, −3.7 percent a year between 1941 and 1944, and −5.1 percent a year between 1941 and 1945. The emphasis on learning by doing has obscured the negative effects of the sudden, radical, and temporary changes in the product mix, the behavioural pathologies accompanying the transition to a shortage economy, and the resource shocks inflicted on the country by the Japanese and Germans. From a long-run perspective, the war can be seen, ironically, as the beginning of the end of US world economic dominance in manufacturing."

Brian R. Cheffins complexifies the narrative of how US antitrust enforcement has evolved in "Getting Antitrust and History in Tune" (*Accounting, Economics, and Law: A Convivium*, published online March 2, 2022, https://www.degruyter.com/document/doi/10.1515/ael-2021-0084/html). From the abstract: "Antitrust is high on the reform agenda at present, associated with calls to 'break up big tech.' Proponents of reform have invoked history with regularity in making their case. They say reform is essential to reverse the baleful influence of the Chicago School of antitrust, which, in their telling, disastrously and abruptly ended in the 1980s a 'golden' era of beneficially lively antitrust enforcement. In fact, antitrust enforcement was, at best, uneven, from the early 20th century through to the end of the 1970s. As for the antitrust 'counter-revolution' of the late 20th century, this was fostered as much by fears of foreign competition and skepticism of government regulation as Chicago School theorizing. The pattern helped to ensure that the counter-revolution was largely sustained through the opening decades of the 21st century."

Paul Krugman delivered the "The Godley–Tobin Memorial Lecture: The Second Coming of Tobinomics " (*Review of Keynesian Economics*, Spring 2023, 11:1, 1–9, https://www.elgaronline.com/view/journals/roke/11/1/article-p1.xml). "James Tobin was, obviously, a Keynesian in the sense that he believed that workers can and do suffer from involuntary unemployment, and that government activism, both monetary and fiscal, is necessary to alleviate this evil. But he wasn't what people used to call a hydraulic Keynesian, someone who imagined that you could analyse the economy by positing mechanical relationships between variables like personal income and consumer spending, leading to fixed, predictable multipliers on policy variables like spending and taxes. . . . Instead, Tobin was also a neoclassical economist. That is, he believed that you get important insights into the economy by thinking of it as an arena in which self-interested individuals interact, and in which the results of those interactions can usefully be understood by comparing equilibria—situations in which no individual has an incentive to change behaviour given the behaviour of other individuals. Neoclassical analysis can be a powerful tool for cutting through the economy's complexity, for clarifying thought. But using it well, especially when you're doing macroeconomics, can be tricky. Why? It's like the old joke about spelling 'Mississippi': the problem is knowing when to stop. What I mean is that it's all too easy to slip into treating maximising behaviour on the part of individuals and equilibrium in the sense of clearing markets not as strategic simplifications but as true descriptions of how the world works . . . So part of the art of producing useful economic models is knowing when and where to place limits on your neoclassicism. And strategic placing of limits is a large part of what Tobinomics is about."

Eric Goldwyn, Alon Levy, Elif Ensari, and Marco Chitti have coauthored "Transit Costs Project: Understanding Transit Infrastructure Costs in American Cities" (New York University, Marron Institute of Urban Management, February 2023, https://transitcosts.com/Final-Report). "Based on our detailed case studies and data collection, we identified three primary factors that comprise total project costs and explain why Phase 1 of New York's Second Avenue Subway is 8 to 12 times more expensive than our composite baseline case. Our baseline case draws on detailed cost data from our Italy, Istanbul, and Sweden cases, and is informed by data from medium-cost Paris and Berlin and low-cost Helsinki and Spain. The New York premium is based primarily on detailed cost data from our Second Avenue Subway case study, and is supported by data from our GLX [Green Line Extension in Boston] case and data from London and Toronto. . . . The good news is that high-cost countries can adopt the practices of low-cost countries and build subways at costs more in line with those of low-cost Scandinavia, Southern Europe, and Turkey. To do this, it requires rethinking design and construction techniques, labor utilization, procurement, agency processes, and the use of private real estate, consultants, and contingencies. If it implements the best practices we detail in the rest of the overview, the highest-cost city in our database, New York, can reduce its construction costs to match those of Italy and match or even do better than Scandinavia."

Symposia

The *Oxford Review of Economic Policy* has published a ten-paper symposium on the "Economics of Pandemic Vaccination." Here, I'll focus on the overview essay, by Scott Duke Kominers and Alex Tabarrok, titled "Vaccines and the Covid-19 pandemic: lessons from failure and success" (Winter 2022, 38:4, pp. 719– 741, https://academic.oup.com/oxrep/issue/38/4). "It is important to note that underinvestment was not simply a US problem—every country underinvested in vaccines. In fact, Operation Warp Speed was by far the largest vaccine investment programme globally, so whatever problems reduced the effectiveness and scale of the US response may have been far larger elsewhere. Global underinvestment in vaccination may in part have been a result of human psychology—voters tend to reward politicians for dealing with emergencies, but not for avoiding them, and in the case of Covid-19, the scales in question may have been especially hard to contemplate. Human psychology may also help to explain why it appears to have been harder to spend trillions on a war against a virus than on wars against other people."

John Baffes and Peter Nagle have edited a four-chapter book on *Commodity Markets: Evolution, Challenges, and Policies* (World Bank, 2022, https://openknowledge.worldbank.org/handle/10986/37404). From the first chapter, "The Evolution of Commmodity Markets over the Past Century," by Baffes and Nagle, together with Wee Chian Koh: "Economic expansion after World War II (WWII), and more recently the emergence of EMDEs [emerging markets and developing economies] as important players in the global economy, has increased commodity demand, especially for energy commodities and metals and minerals. Even though the world's population rose from 2 billion in 1920 to 8 billion in 2020, the production of commodities to feed, clothe, and support the rising population has more than kept pace. Expanding production was possible because of technological innovations, the discovery of new reserves of commodities, and more intensive agricultural production. On the energy front, crude oil became the most important commodity, replacing coal. Known reserves of crude oil and natural gas have increased substantially even as production has risen. . . . Mineral resource development expanded because of advances in technology and new discoveries. Metal production has become more efficient as innovations and productivity improvements became widespread in mining, smelting, and refining. Improved fabrication and new alloys have allowed less metal to be used without loss of strength. Despite radical changes in supply and consumption, metals prices, in real terms, have seen cycles around a quite flat trend over the past century. . . . Food production has increased faster than population, and most of the world's consumers have better access to adequate food supplies today than they did a century ago. This improvement is due to technological advances in the 1900s, especially the Green Revolution. In large part because of increasing productivity, prices of agricultural commodities have experienced a downward trend over the past 100 years."

Regional Development

The United Nations Development Programme has published "Arab Human Development Report 2022: Expanding Opportunities for an Inclusive and Resilient Recovery in the Post-Covid Era (June 2022, https://arab-hdr.org/report/ahdr-2022/). "Few Arab States have competitive private sectors, particularly for tradables, and the economies that are based on oil and gas are subject to highly volatile prices. The productivity of labour, much of it informal, is relatively low. The capacity of government institutions is generally weak. This persistent equilibrium of low growth–low productivity–low employment–low institutional capacity emerges from a social contract based on a deep-seated rentier state that favours the status quo and rejects truly transformative economic reforms. The region's well-known economic fragilities are not destiny, however. They can be corrected with a strong human development approach to tackle the region's long-term structural challenges. Five priorities stand at the top of the agenda. First is to diversify and transform the Arab States region's economies to reduce exposure to commodity cycles and macroeconomic volatility. Second is to boost growth to generate decent jobs for poor people and informal workers and for new entrants to the labour force. Third is to improve the investment climate and level the playing field for businesses, large and small, and investors, domestic and foreign. Fourth is to increase access to finance for women and smaller enterprises. And fifth is to pursue regional economic integration to expand markets and deliver regional public goods."

UNCTAD has published its "Economic Development in Africa 2022" report, "Rethinking the Foundations of Export Diversification in Africa: The Catalytic Role of Business and Financial Services" (July 2022, https://unctad.org/edar2022). "[T]he most relevant variable to promote intra-African bilateral diversification with regard to exporters is a larger share of services value added. By providing business services and market knowledge, information and communications technology (ICT) services facilitate tapping into new markets with new or existing products. Further, transport and distribution services are important across value chains to store and sell products. Access to financial services and research and development are essential to innovation of new products and the continuous improvement of products to survive in markets. Business services can be employed to overcome structural constraints through marketing and consulting to position products on the market. . . . While the world has in recent decades experienced a boom in technology, the use of advanced technologies in the economy remains a challenge for many African countries. Many localities in Africa do not have access to stable Internet connections, in addition to multiple power shortages. Instability in Internet connections slows down services and makes technologies in trade in services less efficient. While digitalization is driving trade in high knowledge-intensive services, Africa remains the least digitalized continent on the planet."

Era Dabla-Norris, Tidiane Kinda, Kaustubh Chahande, Hua Chai, Yadian Chen, Alessia de Stefani, Yosuke Kido, Fan Qi, and Alexandre Sollaci have written "Accelerating Innovation and Digitalization in Asia to Boost Productivity "(International Monetary Fund, January 2023, https://www.imf.org/en/Publications/Departmental-Papers-Policy-Papers/Issues/2023/01/08/Accelerating-Innovation-and-Digitalization-in-Asia-to-Boost-Productivity-523807). From their abstract: "For many Asian countries, the COVID-19 crisis opened deep economic scars, which has led to intensifying pre-pandemic weaknesses—most notably, declining productivity growth. While no panacea will reverse productivity losses, digitalization and innovation can provide a way out. Digitalization can mitigate scarring during downturns—for example, by facilitating virtual education, remote work, and contactless sales—while improving productivity and innovation during expansions. . . . Despite these successes, Asia still faces important divides that prevent it from fully reaping the benefits of innovation-led growth. . . . Within countries, diffusion of innovation from high-performing firms to other firms is limited, including due to constraints in access to finance, management capabilities, and skill gaps in information and communications technologies. Digital gaps and unequal access to digital technologies prevent a sizeable share of firms and workers from reaping the full rewards of participating in the new economy and reaching their full potential."

The Economic Commission for Latin America and the Caribbean has published "A digital path for sustainable development in Latin America and the Caribbean" (November 2022, https://conferenciaelac.cepal.org/8/en/documents/digital-path-sustainable-development-latin-america-and-caribbean). "Today, more than ever before, improvements in inclusion, equality and productivity are associated with the accumulation of new capacities in the area of digital technologies. . . . [I]n a world in which technological progress has accelerated sharply, there is less room for competition based solely on static comparative advantages, such as abundant natural resources or low-skilled labour. To boost economic development, resources need to be reallocated toward innovation- and knowledge-intensive activities; and economies need to diversify into sectors in which both domestic and external demand are growing rapidly. It is undeniable that the digital transformation entails major disruptions that could promote greater inclusion and equality and also foster diversification of the production structure and sustainable productivity growth. Digitalization is affecting all sectors of the economy and society, adding value along the production chain; but the magnitude of the change will depend, largely, on enabling factors such as skills and infrastructure. These technologies have expanded possibilities for advancing towards progressive and inclusive structural change. However, it is also true that the corresponding opportunities are not open to all countries or sectors alike. In fact, rapid digital transformation can become an additional source of social and productive segregation, both within and between countries, if the infrastructure and basic capacities needed to use the technologies appropriately and effectively are not in place. Moreover, success in harnessing the digital revolution depends increasingly on how economies, production sectors, institutions and societies position themselves to absorb and adapt to these changes."

Interviews

Jeff Horwich serves as interlocutor in "Jón Steinsson interview: Forward guidance, the state of macro, and how the economy is like a rumbling volcano" (Federal Reserve Bank of Minneapolis, December 19, 2022, https://www.minneapolisfed.

org/article/2022/jon-steinsson-interview-forward-guidance-the-state-of-macro-and-how-the-economy-is-like-a-rumbling-volcano). "We economists often get criticized for our inability to predict how things are going to turn out. I think people don't fully appreciate the fact that we're trying to predict something that is not only pretty complicated, but we've only seen very few instances of this thing we're trying to predict. In the United States in the post-war period, we've seen maybe a dozen recessions. We've seen inflation really rise three or four times. This is a little bit like a weather forecaster who's trying to predict the weather but has only ever seen 12 storms. . . . And recessions are heterogeneous: The COVID recession is very different from the Great Recession, which is very different from the Volcker recession. There's a similar thing going on in Iceland at the moment, because there's this volcano that is rumbling and actually has erupted twice in the last two years. But this volcano hadn't erupted for 800 years. The volcanologists are on TV every day being asked to predict when the next eruption is going to happen, how long is the eruption going to last, is it going to get bigger, is it going to get smaller. I felt like it was very similar to us economists who are being asked the same questions about events that only happen every decade or so. I think the public is more understanding of the volcanologist than they are of the economist in this respect!"

David A. Price has an interview with "Steven Davis: On remote work, changes in recruiting, and business startups after the pandemic" (*Econ Focus*: Federal Reserve Bank of Richmond, Fourth Quarter 2022 , pp. 22–26, https://www.richmondfed.org/publications/research/econ_focus/2022/q4_interview). "Millions of people left the labor force in spring 2020 when the pandemic struck. . . . It would seem like a simple thing to know exactly how many, but it turns out not to be so easy . . . [T]he data sources that actually track large numbers of people over time in a way that makes it possible to get a precise answer to this question don't become available for two or three years after the fact. And even then, they're hard to access. . . . There are a few things going on, but let me mention two that I think are important. One is that there is increasingly good evidence that out of the tens of millions of people who had COVID-19, a small fraction of them have symptoms that endure for months and months. . . . But let's say you have a hundred million people who had COVID-19—I'm just going to use round numbers here—and 15 percent of them have symptoms that last a long time. The numbers are in that ballpark. Of that 15 percent, let's say a third of them, just to make the arithmetic easy, have pretty serious debilitating conditions like shortness of breath or brain fog, that kind of thing. Now we're talking about 5 million people. Well, you take 5 million people out of the labor force, that's a reduction on the order of 3 percent. That's the long COVID impact on labor force participation, which others have worked on. And then there's long social distancing . . . [S]ome people who used to be in the labor force are now staying out of the labor force because they worry about infection risks associated in the workplace or on the commute to and from work. I think both long COVID and long social distancing are part of the story as to why labor force participation rates haven't recovered fully."

Conversation Starters

Lucian A. Bebchuk and Roberto Tallarita discuss "The Perils and Questionable Promise of ESG-Based Compensation" (*Journal of Corporation Law*, Fall 2022, 37–75, https://jcl.law.uiowa.edu/articles/2023/01/perils-and-questionable-promise-esg-based-compensation). They focus on the 97 US companies in the S&P 100—which together represent over half the total value of the US stock market. "We found that slightly more than half (52.6 percent) of these companies included some ESG [environmental, social, or governance] metrics in their 2020 CEO compensation packages. These metrics focus chiefly on employee composition and employee treatment, as well as customers and the environment, but also, to a much smaller extent, communities and suppliers. ESG metrics are mostly used as performance goals for determining annual cash bonuses. However, most companies do not disclose the weight of ESG goals for overall CEO pay, and those that do disclose it (27.4 percent of the companies with ESG metrics) assign a very modest weight to ESG factors (between less than 1 percent to 12.5 percent, with most companies assigning a weight between 1.5 percent and 3 percent)."

Rachel Silverman Bonnifield and Rory Todd discuss "Opportunities for the G7 to Address the Global Crisis of Lead Poisoning in the 21st Century: A Rapid Stocktaking Report" (Center for Global Development, 2023, https://www.cgdev.org/publication/opportunities-g7-address-global-crisis-lead-poisoning-21st-century-rapid-stocktaking). "Lead poisoning is responsible for an estimated 900,000 deaths per year, more than from malaria (620,000) and nearly as many as from HIV/AIDS (954,000). It affects almost every system of the body, including the gastrointestinal tract, the kidneys, and the reproductive organs, but has particularly adverse effects on cardiovascular health. According to the World Health Organization (WHO), it is responsible for nearly half of all global deaths from known chemical exposures. Despite this massive burden, the greater part of the harm caused by lead may come not through its effects on physical health, but its effect on neurological development in young children. . . . An estimated 800 million children—nearly one in three globally, an estimated 99 percent of whom live in low- and middle-income countries (LMICs)—have blood lead levels (BLL) above 5 micrograms per deciliter (μg/dL), which the WHO uses as a threshold for recommending clinical intervention to mitigate neurotoxic effects." They discuss common vectors of exposure to lead including recycling of lead-acid batteries, along with lead used in spices, paint, cookware, toys, and others.

AMERICAN ECONOMIC ASSOCIATION

AEA 2022 Distinguished Service Award Recipient

Christopher "Kitt" Carpenter
Vanderbilt University

Kitt Carpenter is the inaugural recipient of the AEA's Distinguished Service Award. He is the E. Bronson Ingram University Distinguished Professor of Economics and Health Policy at Vanderbilt University, Director of its Program in Public Policy Studies, and founder and Director of Vanderbilt's LGBTQ+ Policy Lab. He has published widely on the effects of legal same-sex marriage; the causes and consequences of youth substance use; and the effects of public policies on health behaviors, such as bicycle helmet use, seatbelt use, smoking, cancer screening, and vaccination.

About the Award

The annual AEA Distinguished Service Award recognizes volunteer efforts of individuals who have served the economics profession in a significant way. While this award is conferred by the American Economic Association (AEA), it is not expected that service be limited to AEA supported activities.

The award winner is recognized at the AEA Awards Ceremony at the ASSA Annual Meeting.

Information on the nomination and selection process is available at **aeaweb.org/go/dsa**.

Annual Nomination Deadline:
October 1

Committee on Economic Education
AMERICAN ECONOMIC ASSOCIATION
·Established 1955·

AEA Distinguished Economic Education Award

2022 Award Recipient

Charles Holt
University of Virginia

Charles (Charlie) Holt, University of Virginia (UVA), is the inaugural recipient of the AEA Distinguished Economic Education Award. Professor Holt has served on the UVA faculty since 1989 and is currently the A. Willis Robertson Professor of Political Economy in the Batten School of Leadership and Public Policy.

About the Award

The annual AEA Distinguished Economic Education Award acknowledges sustained and impactful contributions to economic education including teaching, the development of curriculum and pedagogy, scholarship of teaching and learning (SoTL), mentoring, and service.

The award is conferred at the Committee on Economic Education's Friends of Economic Education Reception, which is held at the ASSA Annual Meeting.

Nominations

Information about the nomination and selection process is available at **aeaweb.org/go/deea**.

**Annual Nomination Deadline:
October 1**

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. Change of address notice must be received at least six weeks prior to the publication month.

The Journal of

Economic Perspectives

Spring 2023, Volume 37, Number 2

Symposia

Spatial and Urban Economics

Universal Health Insurance

Economics of Mental Health

Article

Features

AMERICAN
ECONOMIC
ASSOCIATION