The Journal of Economic Perspectives

A journal of the American Economic Association

Summer 2019

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), Summer 2019, Vol. 33, No. 3. The JEP is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $22.00, $32.00, or $42.00 depending on income; for an additional $15.00, you can receive this journal in print. The journal is freely available online. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 33 • Number 3 • Summer 2019

Symposia

Markups

Issues in Antitrust

Articles

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—Volume 33, Number 3—Summer 2019—Pages 3–22

Are Price-Cost Markups Rising in the United States? A Discussion of the Evidence

Susanto Basu

Has the US economy entered a second Gilded Age? A pattern of increasing industrial concentration combined with rising inequality in income and wealth may seem to indicate as much. Yet industrial concentration can be interpreted as evidence either of increased market power or of greater competition, where more efficient firms are able to gain market share. Thus, economists have sought to estimate a less ambiguous measure of market power in order to see whether firms are able to exert greater control over the market prices of their outputs. The summary measure that has been the focus of much recent research is the markup of price over marginal cost.

Three main methods have been used to estimate markup trends in the US economy. The first method attempts to estimate economic profits using either aggregate or firm-level data and then, together with an assumption of constant returns to scale, generates an estimate of the size of markups. The second approach estimates a production function for various firms or sectors, based on a variety of inputs. Unlike the first approach, this allows for increasing returns to scale and recovers the markup by applying conditions for cost minimization to the estimated coefficients in the production function. The third method again estimates a production function, typically using firm-level data, but this time recovers the markup from the optimization condition for a single input. This approach again allows for increasing returns to scale, but unlike the first and second methods, it avoids the need to

■ *Susanto Basu is Professor of Economics, Boston College, Chestnut Hill, Massachusetts, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is susanto.basu@bc.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.3 doi=10.1257/jep.33.3.3

estimate the rate of economic profit (or to assume that it is zero). This method estimates the markup as the elasticity of output with respect to the single input, divided by the factor payment to the selected input as a share of the firm's revenue.

I begin with a conceptual overview of these three approaches, which shows how they are connected and provides an intuitive sense of how they allow estimation of markups for the economy as a whole. In the following sections, I examine empirical research that has implemented each of these approaches. I describe in more detail how the research was done and characterize some strengths and weaknesses of all three approaches. I show why estimates of large or steeply rising markups are implausible; for example, several of the prominent estimates suggest that the markup increased far more than would be necessary to explain the decline in labor's share.

The article offers some suggestions for future research on markups. The conclusion, in particular, asks researchers to link microeconomic estimates of markups to aggregate trends in the economy. The recent interest in market power stems in substantial part from the realization that higher markups may tend to depress the demand for factors of production, and thus the prices (incomes) those factors receive. For example, connections have been proposed between higher markups and a lower labor share of income and a lower investment rate. Yet, higher markups should also reduce hiring of workers and, ceteris paribus, raise the inflation rate. Several of the theoretical predictions of what patterns should accompany a substantial rise in markups are not easily verified in recent US data.

Three Methods of Measuring Markups

The markup of price over marginal cost is a basic measure of market power. With perfect competition in the goods market, a profit-maximizing firm will set price equal to marginal cost, and the markup will be equal to one. With imperfect competition, the firm produces at the quantity where marginal revenue equals marginal cost, and price will exceed marginal cost. In seeking to measure markups, an immediate hurdle is how to measure marginal cost—a variable that must be estimated or inferred rather than being directly observed in a market transaction like prices or revenues. Economists specializing in industrial organization have developed ways of estimating markups for particular firms and industries. But the challenge here is to develop measures of average markups for the economy as a whole.

This section describes three theoretical approaches that researchers have used. Those who attempt to estimate markups in a comprehensive manner, for most or all of the US economy, typically use a version of the cost-minimization framework described here.[1] The following three sections then describe in more detail how

[1] De Loecker and Eeckhout (2017) discuss the relationship between this approach and the traditional industrial organization approach to estimating markups using estimates of demand for individual goods or markets.

each approach has been implemented in recent empirical research and discuss some strengths and weaknesses of each approach.

Consider a firm producing output, Y, with a production function, F, that uses capital and labor as inputs, as well as freely available technology, Z: $Y = F(K, L, Z)$. (Researchers using firm- or industry-level data typically add intermediate inputs as a factor of production.) Suppose furthermore that while the firm may have market power in the goods market, it takes the prices of the two factors, the wage, W, and the required return to capital, R, as set in markets outside its control. Then, a profit-maximizing firm will make a cost-minimizing use of labor, which requires that it hire labor until the marginal product of labor equals the markup times the wage:[2]

$$PF_L = \mu W.$$

Here F_L is the marginal product of labor, P is the price of output, and μ is the markup of price over marginal cost. Perfect competition in the goods market corresponds to $\mu = 1$, which yields the familiar condition that an optimizing firm must equate the marginal product of labor to the real wage. However, a firm with market power hires less labor (and thus has a higher marginal product of labor for a given real wage), because it maximizes profit by producing less than the competitive level of output. More market power—a larger markup—corresponds to a lower level of desired output. Naturally, a similar condition applies to the equality between the value marginal product of capital and its rental rate, R, multiplied by μ.[3]

As we will see, the robust cost-minimization conditions alone allow us to *measure* the extent of a firm's market power, even though they do not by themselves answer the interesting question of *why* the firm has market power. Thus, the firm optimality condition above holds regardless of the form of imperfect competition that generates the markup. The firm can be a monopolist or an oligopolist, and it may follow either static or dynamic pricing policies. The firm's optimal choice of the markup is determined by its larger profit-maximization problem, of which cost minimization is only a part. But because we do not need to take a stand on the rest of the firm problem, which can be very complicated, we are able to measure the size of the markup with minimal assumptions.

One conclusion follows from multiplying the condition above by labor input and dividing by output, which has the effect of expressing this relationship in elasticity form:

$$\frac{F_L L}{Y} = \mu \frac{WL}{PY}.$$

[2] The fundamental condition arising from profit maximization is that the firm equates the marginal product of labor valued at marginal cost to the wage. The equation in the text follows from observing that marginal cost by definition equals the ratio of the output price to the markup, which is itself price over marginal cost.

[3] This cost-minimization condition holds even if it is costly for the firm to adjust its capital stock. However, in this case, the rental rate, R, must be redefined to include the marginal adjustment cost of capital and its expected rate of change (for discussion, see Basu, Fernald, and Shapiro 2001).

The left-hand side is the elasticity of output with respect to labor input.[4] The right-hand side is labor's share in revenue, multiplied by the markup. For a given output elasticity—and certainly in the Cobb–Douglas case, where the elasticity is constant—an increase in the markup depresses the share of revenue or national income going to labor.

This implication of higher markups has drawn much interest in recent years, because the labor share of income in the United States has fallen sharply over the decades since 1980. After averaging about 0.64 in pre-1980 data, observations of labor's share have most recently been around 0.58, a sharp decline for what was sometimes called one of the "great ratios" of economic growth. Elsby, Hobijn, and Şahin (2013) attribute around one-third of the measured decline in the share to incorrect measurement of self-employment income, which still leaves an actual decline of four percentage points to be explained.

How much would the markup have to rise to explain this decline in labor's share? If the output elasticity of labor were constant, then the markup would have to increase by a factor of 1.07. That is, if perfect competition ($\mu = 1$) prevailed in 1980, we would require $\mu = 1.07$ now, so price would now be 7 percent higher than marginal cost. As we will see, this implied increase in the markup is modest relative to many of the estimates in the literature. If the markup is to have risen by more, the output elasticity of labor must have *risen* substantially in order to be consistent with the observed change in the labor share.[5] Of course, this back-of-the-envelope exercise attributes all of the decline in the labor share to changing market power. Elsby, Hobijn, and Şahin (2013) attribute much of the change to increased trade competition from globalization, while Karabarbounis and Neiman (2014) attribute it to changes in technology (their explanation implies that the output elasticity of labor should have *fallen* instead of rising).

As this quick calculation demonstrates, inferring changes in markups from changes in observed factor shares requires either an assumption about or, preferably, an estimate of the output elasticity in question. This is the method followed by De Loecker and Eeckhout (2017), which is one of the three approaches to markup estimation discussed at greater length below. They estimate an output elasticity with respect to input econometrically and divide it by the observed revenue share. By the equation above, the ratio gives an estimate of the markup. Repeating this exercise over time, they can also compute a trend in the markup.

The second method I review writes down an equation for efficient use of capital that is parallel to the condition for labor above, and adds the two equations to yield:

$$\frac{F_L L}{Y} + \frac{F_K K}{Y} = \mu \left[\frac{WL}{PY} + \frac{RK}{PY} \right] = \mu \frac{\text{Total cost}}{\text{Revenue}} = \mu(1 - s_\pi).$$

[4] By definition, the elasticity is $\frac{\partial \ln Y}{\partial \ln L}$. Note that $\frac{\partial \ln Y}{\partial \ln L} = \frac{\partial Y / \partial L}{Y/L} = \frac{F_L}{Y/L} = \frac{F_L L}{Y}$.

[5] This observation is due to Brent Neiman.

The left-hand side is the sum of the output elasticities of the production function, which is also the degree of returns to scale. The right-hand side is the markup times the ratio of total cost (including the rental cost of capital) to revenue. Since cost equals revenue minus economic profit, the right-hand side can also be written as the markup times one minus the profit rate s_π, the ratio of profit to revenue.[6]

This equation implies a different method of computing the markup, which is followed by another major strand of the literature discussed below. Suppose one estimates the degree of returns to scale in production, or simply assumes that returns to scale are constant (so the left-hand side equals one). Then one can compute the profit rate and thereby estimate the markup. As the equation above shows, the key to computing the profit rate is to impute a required return to capital, R. Under constant returns and competition, required payments to capital are just total revenue less payments to labor. But with imperfect competition, required payments to capital need to be estimated separately. Once capital payments are estimated and returns to scale are known, one can back out the markup.

This core relationship between markups, economic profit, and economies of scale also clarifies how a firm might have both a markup in excess of one and a near-zero rate of economic profit. This outcome is possible if the firm is producing in the area of increasing returns to scale, where average cost exceeds marginal cost. For example, this would be the case in the classic Chamberlinian model of monopolistic competition, in which long-run profits are zero due to free entry, but firms have market power to set price above marginal cost. Conversely, firms operating with increasing returns to scale—a situation that can arise when marginal costs are low compared with fixed costs—will find that they need to charge a markup above marginal cost to cover their fixed costs, or else they will make losses and go out of business. Of course, in both situations there is a welfare loss arising from the markup even with zero profits, as there is from any wedge (such as a tax) between price and marginal cost. Thus, contrary to suggestions in some papers, the markup is generally a better measure of market power than the profit rate.

Finally, one can derive a third method of estimating the markup by applying the same condition for cost minimization in a different context. This method, due to Hall (1988, 1990), begins by taking a first-order approximation in logs to the production function and taking differences over time of the resulting expression. Letting a lowercase letter represent the natural log of its uppercase counterpart (for example, $y \equiv \ln Y$) and letting a Δ represent a change over time, one gets:

$$\Delta y \simeq \frac{F_L L}{Y} \Delta l + \frac{F_K K}{Y} \Delta k + \Delta z,$$

[6] One can also derive this equation by manipulating the definition of the markup as the ratio of price to marginal cost. Multiply and divide by average cost, then recognize that the ratio of average to marginal cost is the degree of returns to scale for a cost-minimizing firm, while the ratio of price to average cost is also the ratio of revenue to total cost.

where Δz has the interpretation of technical change. Applying the conditions for cost minimization noted above, this equation becomes:

$$\Delta y \simeq \mu\left[\frac{WL}{PY}\Delta l + \frac{RK}{PY}\Delta k\right] + \Delta z.$$

As Hall (1990) emphasized, this approach generalizes Solow's (1957) classic method for calculating the growth rate of technology. If $\mu = 1$, as Solow assumed, then one can obtain a time series for technical change, Δz, as a residual by subtracting share-weighted input growth from output growth. (In Solow's case, required payments to capital are also easily observed as revenue minus labor payments, since there are no profits.) In the case where μ is allowed to exceed one but is unknown, it must be estimated econometrically, using the equation above as an estimating equation, with the unobserved Δz treated as the error term. Since one expects that changes in input usage, the composite right-hand-side variable, would be correlated with the change in technology, Hall uses an instrumental variables technique, where valid instruments must be correlated with input choice (the weighted average of Δk and Δl) but uncorrelated with technical change—loosely speaking, any type of "demand shock." Hall (2018) uses a small modification of this method to estimate markups using recent data.

All three of these methods begin from the assumption that firms minimize costs taking input prices as given. This hypothesis is powerful, but it does not cover all important cases. For example, it assumes that individual firms do not have the power to set wages for their workers. A Council of Economic Advisers (2016) issue brief discussed evidence suggesting, to the contrary, that often firms do indeed have some power to set wages. Qualitatively, the implications of market power discussed above do not change much if firms also have power to set some factor prices. In most cases, such factor market power would also create a wedge between marginal products and factor prices, thus reinforcing the conclusions discussed above for the case of market power in the goods market alone. However, Hall (2018) shows via an insightful example that the quantitative conclusions drawn from applying cost minimization to data to estimate goods-market markups will typically give incorrect results if firms have market power in factor markets as well. Morlacco (2019) presents conditions under which one can reinterpret the evidence for *goods* market power obtained using the method of De Loecker and Eeckhout (2017) noted above as evidence of market power in the *factor* market instead.

Markup Estimates Based on Economic Profits and Constant Returns to Scale

As discussed in the previous section, an assumption of cost minimization makes it possible to derive a relationship between three parameters: returns to scale, the markup, and the rate of economic profit. If returns to scale are assumed to be constant, then calculations of economic profit will allow an estimate of markups.

Barkai (2016) applies this method to US national accounts data and obtains an estimate of the aggregate profit rate, which implies an average economy-wide markup. However, since aggregate time-series data are sparse and explanations for their behavior are typically abundant, Gutiérrez and Philippon (2017a, b) study cross-sectional data at the firm level from Compustat, which provides balance-sheet data on publicly listed US firms. In either case, because the profit rate is typically calculated period by period, this method produces a time series for the implied markup.

Perhaps the main advantage of this approach to markup estimation is that it avoids the need for econometric estimation of production functions, with the attendant difficulties of identification, which are used in the methods that follow. Conversely, the main problem with this approach is that economic profits are notoriously hard to calculate. A typical assumption in this approach is that profits are paid only to owners of capital. This assumption simplifies the computation, because observed payments to labor (and for intermediate inputs in firm- or industry-level data) can be treated as true factor costs, without any profit component. But one still faces the daunting challenge of separating required payments to capital (what the capital would earn on a competitive market) from economic profits, which are really a return to ownership of the firm but are bundled in the data with the implicit rental payments to capital. Efforts to separate the two require the researcher to impute a required return to capital, which when multiplied by the value of the capital stock yields the implicit rental payments.

The required rate of return includes the risk-free real rate, which can be observed from market interest rates on government debt: since 1997, inflation-indexed US government bond yields are available; prior to that, one needs to use nominal yields and subtract an expected inflation rate. It also includes the expected risk premium in excess of the safe rate, which typically must be imputed using an asset-pricing model. Barkai (2016) uses the AAA bond yield as the required return, which includes some compensation for risk. Gutiérrez (2017) explicitly imputes a risk premium, which he adds to a risk-free rate.

Another important component of the rental rate is the economic depreciation rate of the capital stock. Depreciation rates vary widely by type of capital; thus, required returns do as well. For example, Fraumeni (1997, table 3) reports annual depreciation rates of 2–3 percent for business structures, 10–20 percent for most types of business machinery, and 31 percent for office computers. The rate of economic depreciation includes the loss to the owner of capital from physical depreciation—the capital wearing out—as well as the expected capital gain or loss from the change in the resale price of the capital good relative to its purchase price. Most of the large depreciation rate for computers, for example, comes from the decline in the price of a computer over time due to technological progress in the manufacture of new computers, and not from the machine wearing out with use.

The rental cost of capital is then calculated as the sum of the required interest rate and the depreciation rate, multiplied by the market value of the capital stock for each type of capital, summed over all capital types.

While disaggregated stocks of capital are tracked at the level of large industries and the economy as a whole, they generally are not available at a firm level, where firm balance sheets typically report only the book (not market) value of the total capital stock. Furthermore, the national income accounts seek to estimate the rate of economic depreciation, while firm statements report only accounting depreciation. Thus, somewhat counterintuitively, the profit rate might be calculated with less error at an aggregate level, as in Barkai (2016), than at the firm level, as in Gutiérrez and Philippon (2017a, b).

Barkai (2016) calculates the profit rate on value added over the period 1984–2014 with US national accounts data. He finds a much lower profit rate at the start of his sample, 2.2 percent in 1984, than at the end, when it rises to 15.7 percent in 2014. The implied markup ratio μ thus rises from 1.02 to 1.19 over this period.

Gutiérrez and Philippon (2017a) calculate two measures of the profit rate. The first, which they term the "net operating margin," does not subtract the full required return to capital: their implicit rental payment includes depreciation, but not an interest rate. They also compute another markup estimate, based on a full user-cost measure subtracting the required interest payment to capital as well as depreciation. This measure rises by 0.05–0.10 over the period 1980–2015. Interestingly, it is relatively flat until about 2000 and then rises, which matches the timing of the change in labor's share, which is also fairly flat from 1980 to 2000 before declining sharply. Their estimated markup by the end of the sample is about 1.1.

While this estimate appears smaller than Barkai's, it is important to keep in mind that Gutiérrez and Philippon (2017a) are reporting a markup on firm *sales*, which is roughly equal to firm-level gross output, while Barkai reports a markup on *value added*, which summed over firms or industries equals GDP. This important distinction will be discussed further in the next section. For now, it suffices to note that on a common value-added basis, the two end-of-sample estimates are almost identical.

Note that the rise in both markup estimates is noticeably larger than what is implied by the decline in labor's share, as discussed in the previous section. If the markup did indeed rise to about 1.2, then to be consistent with the decline in labor's share corrected for mismeasurement, the output elasticity of labor would have to have risen by about 10 percent over the same period. It is not clear what could have caused such an increase.

A number of refinements to such calculations may be required to calculate the rental cost of capital, and therefore markups, with greater accuracy. Three refinements are worth particular mention. First and most straightforward is to correct the required return for taxes, following the classic method of Hall and Jorgenson (1967).

Second, and much more difficult, is to correct for adjustment costs of capital. This refinement could make use of Hayashi's (1982) neoclassical interpretation of Tobin's (1969) *q* ratio, the value of installed capital relative to the purchase price of new investment goods. The valuation of both the existing capital stock and its expected rate of change should be done using the shadow value of installed capital, marginal *q*, which can differ from the observed purchase price of capital due to adjustment costs. However, these marginal adjustment costs cannot be observed directly and must be estimated econometrically.[7]

Third, the measure of capital could be expanded to include "intangible capital," which seems to be growing in importance in the US economy (for discussion, see Corrado, Hulten, and Sichel 2009). While some forms of intangible capital, primarily software and research and development, are included in the US national accounts, most are not. It is possible that the imputed rentals to capital are too low, because tangible capital is substantially smaller than the true quantity of tangible plus intangible capital. It should be noted that Gutiérrez and Philippon (2017a, b) do incorporate intangible capital into their analysis, using methods in the literature following Corrado, Hulten, and Sichel (2009). However, it is possible that intangible investments grew even faster than the traditional measurements imply, as suggested by McGrattan and Prescott (2010).

In a closely related approach, Karabarbounis and Neiman (2018) also emphasize that the efforts to measure the rental rate of capital may require significant adjustments. They focus on the gap between revenue and imputed total costs, which at the national level is the sum of labor payments and imputed capital rents. They term this gap *factorless income,* because it cannot be attributed easily to either labor or capital, and examine its time-series behavior in aggregate US data from 1960 to 2016. (The method of measuring the markup discussed in this section assumes that "factorless income" represents economic profit resulting from market power, and that it is indeed a return to firm ownership, rather than a required payment to either factor of production.)

It turns out that their measure of factorless income was quite high in the 1960s and early 1970s, then declined, and has been high again since the 1990s. If factorless income is interpreted as profits, then markups must also have been high before 1980. But most of the hypotheses advanced to explain high market power—such as the rise of "superstar firms" (Autor et al. 2017) and a high industrial concentration ratio—fit the recent period but not the pre-1980 period. Similarly, while growth of intangible capital appears able to help explain high levels of factorless income in recent years (if it is interpreted as a return to unmeasured intangibles rather than economic profits), estimates of the quantity of intangible capital typically find that it has been rising steadily over time and was not large before 1970.

Somewhat by default, Karabarbounis and Neiman (2018) suggest that researchers are failing to measure the required rate of return to capital properly. Yet they do not demonstrate that including one or more of the variables typically omitted from the construction of the rental rate can actually account for a substantial

[7] In principle, the value of Tobin's *q* can be inferred from market prices of a firm's debt and equity. This method has two drawbacks. First, it applies only to publicly listed firms. Second, it requires the researcher to assume that asset market valuations reflect only fundamentals at all points in time, a "no-bubble" assumption that may be difficult to justify after the asset market run-ups and crashes observed around the world historically and in the past three decades.

share of the mysterious factorless income. Thus, Karabarbounis and Neiman's (2018) main contribution is to show that plausible changes in market power alone are unlikely to explain the full post–World War II time series of imputed profits/factorless income in aggregate US data.

Estimates Based on Econometric Estimation Using All Inputs

The methods discussed in this section and the next drop the assumption of constant returns to scale. Instead, they rely on estimates of production functions. These production functions may be estimated while imposing the cost-minimization conditions, as in Hall's approach, to give a one-step estimate of the markup. Or the estimated output elasticity for one of the inputs can be compared with that input's revenue share to obtain a two-step estimate, as in the approach of De Loecker and Eeckhout. Both applications also use firm- or industry-level data, where the output concept is gross output and intermediate inputs are an additional factor of production. These changes make no difference to the theory sketched above, but they do change the interpretation of the resulting estimates in an important way.

As noted in the methods section above, the approach pioneered by Hall (1988, 1990) and used recently by Hall (2018)—as well as a large intervening literature!—naturally estimates the markup as a single parameter over the entire sample period. By itself, this method would not provide an estimate of the change in the markup over time, which is the primary focus of the recent literature. Hall (2018) parameterizes each industry-level markup as the sum of a constant and a time trend, and he reports estimates for the weighted average markup at the beginning and end of his sample period, 1988–2015. (As noted above, the method is based on a first-order approximation to the production function, which implies that the output elasticities should be constant over time. To be consistent with the method while allowing a smooth rise in the markup, each input's share should be trending downward at the rate the markup is increasing—an implication that could be checked against the data.)

Hall uses an instrumental variables technique to address the concern that the error term is endogenous. Specifically, Hall uses four categories of military expenditures and the price of West Texas intermediate crude oil as instrumental variables that are arguably uncorrelated with technical change. Hall (2018) applies this technique to US data for 60 industries. Most of these industries are at the North American Industry Classification System (NAICS) two- or three-digit level of aggregation; some examples of large industries in the dataset include retail trade, wholesale trade, and construction.

A significant advantage of Hall's (2018) method is that it does not constrain returns to scale to be constant. The disadvantage is that it requires strictly more information, as well as econometric estimation with instrumental variables.

Notice that Hall's (2018) method still requires its user to compute the quantity of profits, because in his one-step production-function approach, the shares are the *cost* of each input divided by revenue. The typical assumption is that profits are received only by capital, so one needs to impute the rental rate of capital, as in the papers discussed in the previous section. Hall acknowledges this issue. But in practice, he follows the construction of the KLEMS dataset from the US Bureau of Labor Statistics, which offers sectoral data on output, as well as inputs and shares of capital (K), labor (L), energy (E), materials (M), and purchased business services (S). The shares are constructed assuming that total cost equals total revenue, which of course is correct only if economic profits are zero. If profit rates are zero, then the estimates that Hall presents as markups are also estimates of the degree of returns to scale, as shown above.

Hall (2018) estimates that the weighted average industry markup is about 1.3 in 2015, and that industry markups fall in the range of 1.0–1.8. (Some 30 percent of the point estimates are below one but are constrained to equal one on economic grounds, because firms would never systematically price output below marginal production cost.) The time trend is estimated to be positive, implying that markups have been rising over time, although the estimate is not statistically larger than zero at conventional levels of significance.

It may appear that Hall's (2018) estimate of average markup is only slightly larger than Barkai's (2016) estimate. This conclusion would be incorrect. Hall is using industry data and estimating a markup on gross output, while Barkai is estimating a markup on value added. A markup on gross output leads to a larger markup on value added when one takes into account the fact that firms use intermediate goods in production (Rotemberg and Woodford 1995; Basu and Fernald 2002). The intuition is that there is a "double-marginalization" phenomenon—firms sell some of their output for use as intermediate goods, which are bought by other firms that levy an additional markup on top of the markup they paid for their intermediates, and so on. Assuming an intermediate input share of 0.50, approximately the average value for the US economy over a long period of time, a markup of 1.3 on gross production translates to a markup on value added of 1.9, far larger than Barkai's estimate.[8]

One way to interpret this estimate, consistent with Hall's (2018) implicit assumption of zero economic profit, is that the production function for GDP using just capital and labor inputs must have returns to scale equal in size to the value-added markup, namely 1.9. To understand the implications of such a large degree of increasing returns in the aggregate production function, consider aggregate US data for 2015 as reported by the Bureau of Labor Statistics (2019). The BLS reports that private nonfarm business sector value-added output grew 3.5 percent in 2015, while the weighted average of capital and labor input in that sector grew 2.7 percent in the same year. For this growth in output and inputs to be consistent with returns to scale of 1.9 requires that true technological progress must have been *negative*

[8] The relationship between the two markup concepts is $\mu = \dfrac{\mu^G(1 - s^M)}{1 - \mu^G s^M}$, where μ is the markup on value added, μ^G is the markup on gross output, and s^M is the intermediate input share of revenue.

1.6 percent! Such a high rate of technological regress for the US economy as a whole seems quite implausible, casting doubt on the high returns to scale implied by Hall's estimate of the average markup in 2015.

Why might Hall (2018) be estimating markups/returns to scale that are too large? The key concerns regarding his procedure are those that arise whenever one attempts to estimate production functions in differences—that is, looking at change in output and changes in inputs—which is the core of Hall's method. First, consistent estimation of scale economies requires that we measure the real quantities of all the inputs correctly, as opposed to just their nominal payments, which is all that is required when computing profit rates. (Strictly speaking, the estimation is consistent if any measurement error in the inputs is uncorrelated with the instruments.) Second, when variables are measured in growth rates, the resulting correlation tends to emphasize high-frequency variation in output and inputs. A substantial macro literature has emphasized that actual capital and labor inputs vary at high frequencies in ways that are not recorded in the conventional production data that Hall uses (Bils and Cho 1994; Burnside, Eichenbaum, and Rebelo 1996; Basu, Fernald, and Kimball 2006). For example, firms may vary the workweek of capital by changing the number of shifts used to produce output, thus changing the true capital service input but without a change in the observed capital stock. Firms appear to vary capital's workweek and labor effort as they change their rate of production, in response to both demand and technology shocks. This unmeasured variation in utilization will probably lead to an upward bias in the estimated markup, and using demand instruments will not solve this problem.

As an example of how such considerations can affect the results, Basu, Fernald, and Kimball (2006) also use annual industry-level production data and a procedure similar to Hall's (2018), but with an additional control for variations in the intensity of factor usage. Their data cover the period 1949–1996, so they do not examine the past two decades (although the sample does cover the early post–World War II period, when there is also evidence of high profit rates/markups). Controlling for variable factor utilization, Basu, Fernald, and Kimball find few industry markup estimates that are greater than one at conventional levels of statistical significance. The clearest evidence of positive markups is in durables manufacturing, but even there the median industry markup is just 1.07 (on gross output).[9] Outside of durables manufacturing, only one industry (chemicals) is estimated to have a markup significantly larger than one. On the other hand, the estimated controls for utilization are positive and highly significant for both durables and nondurables manufacturing industries, and large although statistically insignificant outside manufacturing.

Thus, future research using Hall's (2018) method might proceed in several ways. First, when using this approach, it might be useful to apply similar utilization controls to see whether, by reducing the effect of short-term variations in

[9] Basu, Fernald, and Kimball (2006) present their results as estimates of returns to scale, but since they operate under the same zero-profit assumption as Hall, their estimates can be interpreted equally well as markups.

unmeasured inputs, one also reduces the markup estimates obtained. Second, it would be useful to investigate further the interpretation of the estimated μ, and the extent to which it is capturing economies of scale, economic profits, or some mixture of the two. Finally, because this approach requires a calculation of profit, all of the questions raised in the previous section about measuring the rental cost of capital apply here as well.

Econometric Estimation Using a Single Input

The theory presented in the first main section of this paper established a framework in which the markup must be the same for each input (because marginal cost must be the same along every margin). Thus, it should be possible to compute the markup on the basis of only a single input to production, not many inputs. Moreover, if one chooses an input that does not receive pure profits, then the issues of measuring required returns to capital and the profit rate do not arise. The single-input method would be an ideal one to apply to data on intermediate inputs, which probably do not share in pure profits and are measured with the least error due to utilization.[10]

De Loecker and Eeckhout (2017) take this single-input approach using balance-sheet data on publicly listed firms from Compustat. In a later version of this paper, De Loecker, Eeckhout, and Unger (2018) also use firm-level data from the US Census, which is a better source for production data and provides information on firms that are privately held as well. As of this writing, their results using these data had not been cleared for disclosure by the Census Bureau, and hence the discussion in this paper is based on the results using only Compustat data.

These authors use firms' expenditures in their accounting reports on a composite input termed cost of goods sold (COGS), which consists of most intermediate goods and a subset of labor input. They take COGS and the fixed capital stock, K, as their two inputs to production at the firm level. By hypothesis no profits are paid to COGS, so they can construct the share of this factor's cost to total revenue simply from reported data. Like many authors in the industrial organization literature, they use a variant of the technique introduced by Olley and Pakes (1996) to estimate a Cobb–Douglas production function without imposing constant returns to scale and obtain the relevant coefficient estimate. Then, they can divide the estimated output elasticity of COGS by its observed revenue share to calculate the markup. Using the cross-section dimension of their data, they are able to estimate

[10] Indeed, Dobbelaere and Mairesse (2013) apply a similar idea to firm-level data from France to allow for the possibility that both labor and capital bargain over the profits generated by markups. They estimate the markup from the intermediate input margin only, and estimate the bargaining power of capital and labor from the other margins. Their technique also suggests a possible method to allow for monopsony and monopoly power in the same estimation framework. Unfortunately, the Compustat data do not allow for this attractive approach, because most firms do not report separate expenditures on labor and intermediate inputs.

a different output elasticity for each factor for all years, and thus a time series for the markup. (Note that while the output elasticity is constrained to be equal across firms at a point in time, the revenue share and hence the markup estimate vary across firms and over time.)

In their headline estimates, De Loecker, Eeckhout, and Unger (2018) report that the weighted average of the markup ratio at the firm level rises from 1.21 in 1980 to 1.61 in 2016, with most of the increase taking place in the 1980s and 1990s.[11] (Note that the timing of the estimated markup change does not match particularly well the timing of the decline in labor's share of national income, which drops sharply starting in the 2000s but is fairly stable earlier.) The authors emphasize that the trend in the average markup is being driven by increased heterogeneity at the firm level. They plot the distribution of their estimated markups across firms in 1980 and 2016. Both distributions have a mode that is very similar, about 1.3, but the distribution in 2016 has a higher standard deviation and a thicker right tail. This increase in density in the right tail leads to the much higher estimate of the average markup by the end of their sample.

Markups as large as those reported by De Loecker, Eeckhout, and Unger (2018) at the end of their sample have some implausible implications. For example, because the authors report that average returns to scale are about 1.05 in 2016, but that the average markup is 1.61, the relationship given earlier between markups, returns to scale, and economic profits suggests that the average economic profit rate must be extremely high, on the order of 35 percent of firm sales. Since sales on average are about twice as large as value added, this calculation suggests that about 70 percent of GDP is pure economic profit! Profit rates of this size are too large to be credible. Also, because labor receives slightly less than 60 percent of GDP as compensation, and the authors assume that none of that payment is profit, economic profits of this size would mean that there is not enough output to pay both labor and profit, let alone any required return to capital.

Another implausible implication arises because, in keeping with the rest of the recent literature, De Loecker, Eeckhout, and Unger (2018) assume that profits are paid only to owners of capital. On average across US industries, firms spend about 50 percent of their revenues on intermediate goods and 30 percent on labor. With these shares and a markup of 1.61, the output elasticity of labor and intermediates must sum to about 1.3. But since they estimate average returns to scale of only 1.05, the implied output elasticity of capital must be on the order of –0.25.[12] It seems very

[11] These are the results from what De Loecker, Eeckhout, and Unger (2018) term the "traditional" production function, which they denote PF1. They also report results for an alternative specification, PF2, in which markups rise from about 1 in 1980 to 1.32 in 2016. However, the paper stresses the results from PF1, which are the only ones mentioned in the abstract and are presented first in the introduction. Thus, the discussion here focuses on the PF1 results, while noting the PF2 estimates when they differ significantly.

[12] On its face, the finding by De Loecker, Eeckhout, and Unger (2018) that returns to scale have not risen over time does not support one common hypothesis for explaining the rise of markups. A standard implication of a "knowledge economy" is that production is characterized by large fixed or sunk costs but

unlikely that firms spend billions of dollars on investment to accumulate capital that will reduce their output. Furthermore, the implied negative capital elasticity is inconsistent with De Loecker, Eeckhout, and Unger's own production-function estimate of the output elasticity of capital, which averages approximately 0.2. This inconsistency between the implied and the directly estimated capital elasticities is an indication that the estimation is producing problematic results.

Finally, if the average markup of 1.61 is put on a value-added basis as before, the implied markup on GDP is in excess of 4, enormously higher than standard estimates in the macro literature, which typically estimates values between 1.1 and 1.4 (for example, Basu and Fernald 1997; Christiano, Eichenbaum, and Evans 2005).

The preceding calculations have taken the average markup reported by De Loecker, Eeckhout, and Unger and examined its implications as if it were the markup of the representative firm in the economy. But given their finding of extreme heterogeneity in markups, these calculations should be performed at the firm level. The authors can easily perform the first two calculations, of profit rates and the implied output elasticity of capital, at the firm level using their existing estimates, with the results reported as distributions. The third calculation, of the value-added markup, is also easily done, but the data may not be available for every firm. The reason is that many firms in Compustat do not report the data necessary to construct an intermediate input share at the firm level, which is required for the conversion. Given the great interest that this paper has generated, many economists would be keen to see these additional results.

The method used by De Loecker, Eeckhout, and Unger (2018) seems excellent in principle, so why is it leading to implausible estimates of markup levels and trends? As a matter of accounting, the sharp rise in the markup the authors estimate could be driven by either a rising output elasticity of their key input measure, cost of goods sold, or a decline in its share over time. In fact, the output elasticity is estimated to be nearly constant over the sample period, so the rise in the estimated markup is being driven completely by the decline in the share.

This fact should motivate us to think harder about this measure of inputs. Cost of goods sold is an accounting concept used to value changes in inventory holdings for firms that produce to stock. But such industries—agriculture, mining, and manufacturing—produce only about 16 percent of private-sector output in the US economy (for data from 2016, see table 5 in Bureau of Economic Analysis 2018). The concept is much less meaningful when applied to the service industries that produce twice as large a share of US GDP, such as finance and insurance, health care, education, and professional and business services.[13] For some intuition behind this issue, one might try to describe how the reported COGS for a few large, publicly

low marginal cost. The classic example is software, which can take huge resources to develop but then can be replicated at essentially zero cost.

[13] In service industries, cost of goods sold is often renamed "cost of revenue," but the two are conceptually similar.

listed service companies (like Facebook, Goldman Sachs, or a for-profit hospital chain such as HCA Healthcare) is meaningful in an economic sense.

In Compustat, firms actually report two measures of operating (noncapital) expenses. One is cost of goods sold; the other is selling, general, and administrative expenses (SGA). If an increasingly larger share of the inputs that used to be classified as COGS is now recorded as being part of SGA, then such mismeasurement could explain why the COGS share of firm revenue is falling over time, and consequently why markups are estimated as rising over time. Traina (2018) redoes the procedure using the sum of COGS and SGA as the measure of noncapital input and finds no evidence that markups have increased over time.

Some of the controversy following Traina's (2018) paper has focused on the extent to which cost of goods sold can be interpreted as a variable cost and selling, general, and administrative expenses as a fixed cost. (Here the word "fixed" is used to mean overhead inputs, ones that do not vary with the amount of output produced, at least locally, as opposed to inputs that are quasi-fixed, meaning costly to adjust.) The controversy is misplaced, because the underlying theory does not require that *all* of the input on the examined margin be variable. It requires only that there be *some* variable inputs in the input bundle under consideration, and that the bundle be defined consistently over time. If overhead inputs are a higher share of total inputs, the estimated output elasticity of that input bundle will be larger—appropriately so, since overhead inputs are one important source of increasing returns to scale.

Given that there is no harm in deriving the markup using an input aggregate that includes some overhead inputs, it is safer to use a more comprehensive input measure. For example, by convention, payments to salaried workers are classified as selling, general, and administrative expenses, while variable (hourly and commission) labor payments are in cost of goods sold. If there has been a general change in compensation practices for workers fulfilling the same function, shifting them from being classified as part of COGS to being included in SGA, then COGS could not be used to compute the trend in the markup, but the sum of COGS and SGA could be used for this purpose.[14]

This single-input approach to estimating markups would be best executed using a single, distinct measure of physical input, such as production-worker labor hours or purchased energy. However, such data are not available for a large fraction of firms in the Compustat dataset. The next-best choice would be to use a comprehensive measure of composite inputs that includes most or all variable inputs and some overhead inputs, but where the measure of inputs is sufficiently comprehensive that we can be confident it is defined consistently over time.

[14] Outsourcing work would reduce the labor component of cost of goods sold but increase the intermediate input component, and it should not change the total appreciably. This hypothesis is consistent with the findings of Meixell, Kenyon, and Westfall (2014).

Directions for Future Research: Reconciling Micro Estimates and Macro Facts

Two of the three approaches to markup estimation reviewed above yield end-of-sample estimates of the average value-added markup that are too large to be credible. By the underlying logic of cost minimization, high markups must be matched either by equally high returns to scale or by large rates of pure economic profit. Yet returns to scale of the magnitude required would imply large rates of technological regress for the US economy, while the implied rates of economic profit would displace all of the required payments to capital, as well as some of the observed payments to labor. In both cases, the larger estimates of increases in the markup greatly overshoot what is required to explain the decline in labor's share: the true puzzle becomes why labor's share has not fallen far more! Only the approach based on constant returns and computed profit rates leads to estimates of the level and change in the markup that might be consistent with the observed decline in labor's share, and even these estimates are on the high side.

It is worth reviewing some other macro implications of higher markups to contrast them in an informal way with recent data for the US economy. According to standard models, higher markups should reduce the demand for inputs of labor and capital, leading for example to weak growth in jobs and wages, and should raise inflation relative to a welfare-theoretic measure of slack in the economy, because higher markups act as "cost-push" shocks to the Phillips curve. In classic endogenous-growth models, higher markups should also spur innovation, as firms compete more fiercely to displace incumbents from profitable markets, leading to higher rates of productivity growth.[15]

With one partial exception, none of these predictions is borne out in recent US data. The labor market is extremely tight, inflation is subdued, and productivity growth has been weak. (The second fact is particularly striking in light of the first and third.) Even wage growth has been stronger than one would conclude from standard data: Daly and Pedtke (2018) find that median weekly earnings adjusted for labor-force composition have grown 1.5 to 2 percentage points faster per year over the period 2013–2017 than the unadjusted data suggest. An adjustment of this magnitude nearly doubles the growth rate of median weekly earnings. Because rising markups should depress labor demand, it is difficult to reconcile the hypothesis of rising markups with strong growth in both the quantity and the price of labor input over the past several years.

One piece of macro evidence that does go in the direction that the rising markup hypothesis predicts is low business investment. Gutiérrez and Philippon (2017a, b) argue that total investment in both tangible and intangible capital has

[15] This prediction is more nuanced in recent models. Aghion and Griffith (2005) show that the relationship between productivity growth and the markup may have an inverted-U shape, with high markups reducing growth. Aghion et al. (2019) suggest that high markups may lead to first higher and then lower rates of total factor productivity growth.

been weak in recent years, particularly when conditioned on Tobin's *q* ratio, a common measure of fundamentals. Higher markups leading to a positive profit rate can indeed make the average (asset-market) *q* high, while making marginal *q*, which determines investment, low. (Markups by themselves are neither sufficient nor necessary to break the link between average and marginal *q*; a positive economic profit rate is the key wedge.)

Another piece of evidence that may be consistent with a rising markup is a low natural rate of interest. The link between the two is that a higher markup is supposed to create expectations of declining consumption growth due to low demand for capital and labor, which in turn should pull down the interest rate. With the two factor markets for capital and labor seemingly moving in opposite directions, it is not obvious that the markup channel is at work. Carvalho, Ferrero, and Nechio (2016) and Gagnon, Johannsen, and Lopez-Salido (2016) show that demographic forces likely explain much of the decline in long-term real rates.

Thus, future research needs to address two puzzles. The first is why most markup estimates based on micro data are implausibly large and grow too fast in relation to the macro facts to be explained. The second is why most macro data appear to indicate that markups are low and stable, but the investment rate is sending a different signal. A full understanding of markup trends and their economic effects requires an explanation of these two issues.

■ *I thank the editors for useful suggestions that have improved the paper, and Jan Eeckhout, John Fernald, Germán Gutiérrez, Robert Hall, Brent Neiman, Thomas Philippon, Valerie Ramey, Fabio Schiantarelli, Chad Syverson, and J. Christina Wang for helpful comments and discussions.*

References

Aghion, Philippe, Antonin Bergeaud, Timo Boppart, Peter J. Klenow, and Huiyu Li. 2019. "A Theory of Falling Growth and Rising Rents." Federal Reserve Bank of San Francisco Working Paper 2019-11.

Aghion, Philippe, and Rachel Griffith. 2005. *Competition and Growth: Reconciling Theory and Evidence.* Cambridge, MA: MIT Press.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2017. "The Fall of the Labor Share and the Rise of Superstar Firms." NBER Working Paper 23396.

Barkai, Simcha. 2016. "Declining Labor and Capital Shares." Stigler Center New Working Paper Series 2.

Basu, Susanto, and John G. Fernald. 1997. "Returns to Scale in U.S. Production: Estimates and Implications." *Journal of Political Economy* 105(2): 249–83.

Basu, Susanto, and John G. Fernald. 2002. "Aggregate Productivity and Aggregate Technology." *European Economic Review* 46(6): 963–91.

Basu, Susanto, John G. Fernald, and Miles S. Kimball. 2006. "Are Technology Improvements

Contractionary?" *American Economic Review* 96(5): 1418–48.

Basu, Susanto, John G. Fernald, and Matthew D. Shapiro. 2001. "Productivity Growth in the 1990s: Technology, Utilization, or Adjustment?" *Carnegie-Rochester Conference Series on Public Policy* 55(1): 117–65.

Bils, Mark, and Jang-Ok Cho. 1994. "Cyclical Factor Utilization." *Journal of Monetary Economics* 33(2): 319–54.

Bureau of Economic Analysis. 2018. "Gross Domestic Product by Industry: First Quarter 2018." Bureau of Economic Analysis News Release, July 20, 2018. https://www.bea.gov/system/files/2018-07/gdpind118_3.pdf.

Bureau of Labor Statistics. 2019. "Private Business and Private Nonfarm Business Multifactor Productivity Tables." March 20, 2019. https://www.bls.gov/mfp/special_requests/mfptable.xlsx.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo. 1996. "Capital Utilization and Returns to Scale," in *NBER Macroeconomics Annual 1995*, vol. 10, edited by Ben S. Bernanke and Julio J. Rotemberg, 67–124. Cambridge, MA: MIT Press.

Carvalho, Carlos, Andrea Ferrero, and Fernanda Nechio. 2016. "Demographics and Real Interest Rates: Inspecting the Mechanism." *European Economic Review* 88(1): 208–26.

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

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2009. "Intangible Capital and U.S. Economic Growth." *Review of Income and Wealth* 55(3): 661–85.

Council of Economic Advisers. 2016. "Labor Market Monopsony: Trends, Consequences and Policy Responses." Council of Economic Advisers Issue Brief, October 2016.

Daly, Mary C., and Joseph H. Pedtke. 2018. "Revisiting Wage Growth." Federal Reserve Bank of San Francisco Economic Letter.

De Loecker, Jan, and Jan Eeckhout. 2017. "The Rise of Market Power and the Macroeconomic Implications." NBER Working Paper 23687.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2018. "The Rise of Market Power and the Macroeconomic Implications." Unpublished, September 14, 2018.

Dobbelaere, Sabien, and Jacques Mairesse. 2013. "Panel Data Estimates of the Production Function and Product and Labor Market Imperfections." *Journal of Applied Econometrics* 28(1): 1–46.

Elsby, Michael W. L., Bart Hobijn, and Ayşegül Şahin. 2013. "The Decline of the U.S. Labor Share." *Brookings Papers on Economic Activity*, Fall, 1–52.

Fraumeni, Barbara M. 1997. "The Measurement of Depreciation in the U.S. National Income and Product Accounts." *Survey of Current Business* 77(7): 7–23. https://apps.bea.gov/scb/account_articles/national/0797fr/maintext.htm.

Gagnon, Etienne, Benjamin K. Johannsen, and David Lopez-Salido. 2016. "Understanding the New Normal: The Role of Demographics." Finance and Economics Discussion Series 2016-080.

Gutiérrez, Germán. 2017. "Investigating Global Labor and Profit Shares." Unpublished, October 2017.

Gutiérrez, Germán, and Thomas Philippon. 2017a. "Declining Competition and Investment in the U.S." NBER Working Paper 23583.

Gutiérrez, Germán, and Thomas Philippon. 2017b. "Investmentless Growth: An Empirical Investigation." *Brookings Papers on Economic Activity*, Fall, 89–190.

Hall, Robert E. 1988. "The Relation between Price and Marginal Cost in U.S. Industry." *Journal of Political Economy* 96(5): 921–47.

Hall, Robert E. 1990. "Invariance Properties of Solow's Productivity Residual," in *Growth/Productivity/Unemployment: Essays to Celebrate Bob Solow's Birthday*, edited by Peter A. Diamond, 71–112. Cambridge, MA: MIT Press.

Hall, Robert E. 2018. "New Evidence on Market Power, Profit, Concentration, and the Role of Mega-Firms in the US Economy." Unpublished, September 23, 2018.

Hall, Robert E., and Dale W. Jorgenson. 1967. "Tax Policy and Investment Behavior." *American Economic Review* 57(3): 391–414.

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

Karabarbounis, Loukas, and Brent Neiman. 2014. "The Global Decline of the Labor Share." *Quarterly Journal of Economics* 129(1): 61–103.

Karabarbounis, Loukas, and Brent Neiman. 2018. "Accounting for Factorless Income." NBER Working Paper 24404.

McGrattan, Ellen R., and Edward C. Prescott. 2010. "Unmeasured Investment and the Puzzling US Boom in the 1990s." *American Economic Journal: Macroeconomics* 2(4): 88–123.

Meixell, Mary J., George N. Kenyon, and Peter H. Westfall. 2014. "The Effects of Production Outsourcing on Factory Cost Performance: An Empirical Study." *Journal of Manufacturing Technology Management* 25(6): 750–74.

Morlacco, Monica. 2019. "Market Power in Input Markets: Theory and Evidence from French Manufacturing." Unpublished, March 20, 2019.

Olley, G. Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica* 64(6): 1263–97.

Rotemberg, Julio J., and Michael Woodford. 1995. "Dynamic General Equilibrium Models with Imperfectly Competitive Product Markets," in *Frontiers of Business Cycle Research*, edited by Thomas F. Cooley, 243–93. Princeton, NJ: Princeton University Press.

Solow, Robert M. 1957. "Technical Change and the Aggregate Production Function." *Review of Economics and Statistics* 39(3): 312–20.

Tobin, James. 1969. "A General Equilibrium Approach to Monetary Theory." *Journal of Money, Credit and Banking* 1(1): 15–29.

Traina, James. 2018. "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements." Stigler Center New Working Paper Series 17.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 23–43

Macroeconomics and Market Power: Context, Implications, and Open Questions

Chad Syverson

Prior research on market power had largely been the domain of microeconomists, who focused their analytical microscopes on individual industries or markets. Decades of microeconomic study have built a knowledge base, formed modeling conventions, and standardized empirical practices. However, a robust debate has erupted about whether the influence of monopoly power has grown beyond its traditionally studied microeconomic realm of the single industry or market and into the economy overall. Empirical investigations have found broad growth in measured profit rates, price-cost margins, and market concentration since at least as far back as 2000, if not earlier. Those upward shifts have been accompanied by drops in measured investment rates, firm entry rates, and labor's share of income. If average levels of market power have indeed grown across the board, this is likely to degrade key metrics of economy-wide well-being, including investment, innovation, total output, and the distribution of income. A related debate, though one in which economists have played a lesser role, involves the consequences of potential broad-based concentration of not just economic but also political and cultural power (for example, Khan 2017).

This article assesses the macroeconomic market power research. I write as someone who has primarily studied market power in microeconomic frameworks

■ *Chad Syverson is Eli B. and Harriet B. Williams Professor of Economics, University of Chicago Booth School of Business, Chicago, Illinois. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is chad.syverson@chicagobooth.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.23 doi=10.1257/jep.33.3.23

but who has also done some macro-oriented research on topics such as aggregate productivity trends, albeit not dealing with market power per se from a macro perspective. For various reasons, the recent macro-oriented work has often departed somewhat from the established practices in micro analysis of market power. Part of this shift is surely tied to the obvious difference in the scope of analysis; things that can be done relatively straightforwardly for an individual market are not so easy to do at the economy-wide level. But there are other differences, too.

I will look at the issue of market power from a number of different perspectives. I begin with a theoretical comparison of defining market power in formal terms as a markup of price over marginal cost on the one hand and the often-used approach of using concentration to measure market power on the other. I then discuss how a prominent strand of macro market power research has used accounting data to estimate markups. I look at how markups are necessarily related to prices, costs, scale elasticities, and profits and point out seeming inconsistencies among the empirical estimates of these values in the literature. I then look at some of the research that has linked a rise in market power to lower levels of investment and a lower labor share of income. Throughout this discussion, I characterize the congruencies and incongruencies between macro evidence and micro views of market power and, when they do not perfectly overlap, explain the open questions that need to be answered to make the connection complete. I hope in this article to pull the two bodies of work somewhat closer together.[1]

To preview my conclusion, I believe the macro market literature has established and collected an important and provocative set of facts, some developed by this literature and some built closely upon previous work. The literature has done a service by drawing plausible connections among these facts and showing how they might be tied to increases in the average level of market power. However, I believe the case for large and general increases in market power is not yet dispositive. There are empirical holes to be filled and plausible alternative stories (some with evidence of their own in their favor) that would first need to be rejected. To be clear, this is not to say that I believe the case for market power has failed or ought to be rejected. It remains a leading candidate explanation for several trends in the data. Rather, to my mind there remains considerable empirical uncertainty around the existence and magnitude of any across-the-board increase in market power in the economy. Thus, I finish the discussion by addressing what holes in the existing literature I would like to see filled.

[1] I focus in this article on research that involves broad-scope empirical examinations of market power and its effects. There continues to be a lot of work examining market power in specific industries or markets. Some have argued for broader implications to be drawn from these market-specific studies. See, for example, Shapiro (2018 and in this issue) or the reaction of Scott Morton and Hovenkamp (2018) to studies like Azar, Schmalz, and Tecu (2018). There does not yet seem to be a consensus about whether market-specific studies themselves, alone or in their collective weight, have macroeconomic implications, but they can exposit the potential mechanisms at work.

Market Power, Markups, and Concentration

The literal textbook definition of market power is a firm having the ability to influence the price at which it sells its product(s) (for example, Pindyck and Rubinfeld 2012; Goolsbee, Levitt, and Syverson 2016). In other words, if a firm does not face a perfectly elastic residual demand curve, it has market power. A connotation of this definition, sometimes left implicit, is that the firm uses this ability to hold the price above marginal cost.[2]

Using this definition, the *magnitude* of market power is tied to the size of the gap between price and marginal cost at the firm's profit-maximizing level of output. The size of this gap—typically called the "markup" when expressed multiplicatively and the "margin" when expressed as a difference, though there is some variation in usage—depends on the shape of the firm's residual demand curve. Steeper inverse demand raises the profit-maximizing margin and implies more market power.

Markups are difficult to measure directly. They require information not just on prices but on hard-to-observe marginal costs. As a result, there are many informal definitions of "market power" and associated metrics used in popular economic writing, and sometimes in economic research as well. Examples include the number of competitors (actual or potential), profit rates, and costs of market entry. Each of these alternatives has its merits, but each is also one step removed from actual pricing power, which in turn can lend itself to shortcomings and ambiguities in practice.

In the recent macroeconomic market power literature, the most frequently used measure of market power is concentration. Measures of market concentration summarize the share of market or industry activity accounted for by large firms. The two most common are the Herfindahl–Hirschman index, which is the sum of firms' squared market shares, and C_n, which is the combined market share of the largest n firms.

An advantage of concentration as an empirical tool for studying market power is that it requires data only on revenues and thus is often relatively easy to compute. The corresponding disadvantage is that concentration is about relative revenue and thus includes no information about costs or profits. For example, a monopolistically competitive market can be very unconcentrated and display near-zero levels of economic profit—indeed, monopolistic competition is defined by the atomistic nature of firms combined with possibilities for entry and exit—but firms in such a market can still have very inelastic residual demand curves and hence a lot of market power.

Concentration also necessarily requires a market definition, which is often a point of contention. The macro literature has in some cases measured the extent of concentration within broad industry groupings, which raises the possibility that

[2] I focus here on market power in output markets. Of course, firms can also exercise market power in the markets for their inputs, including labor; for discussions, see Manning (2003), Azar, Marinescu, and Steinbaum (2017), Benmelech, Bergman, and Kim (2018), and Krueger (2018).

increases in concentration in narrower and more relevant markets may be invisible in the broader measures. Moreover, national concentration measures can be particularly misleading for geographically localized markets. For example, a chain restaurant building stores in a number of local markets would tend to increase measures of concentration computed at the national level, even if it reduced concentration in the economically relevant local markets. Rossi-Hansberg, Sarte, and Trachter (2018) find evidence suggesting that a "national concentration, local de-concentration" pattern is occurring in a number of industries.

Perhaps the deepest conceptual problem with concentration as a measure of market power is that it is an outcome, not an immutable core determinant of how competitive an industry or market is. The nature and intensity of industry competition combine with other supply and demand primitives to determine equilibrium concentration. However, the conditions of competition drive concentration, not vice versa.

As a result, concentration is worse than just a noisy barometer of market power. Instead, we cannot even generally know which way the barometer is oriented. Even if researchers agree on a definition of the market, concentration can be associated with either less or more competition.

Research that uses concentration as a measure of market power is implicitly relying on the mechanics of the standard Cournot oligopoly model, in which a number of firms with possibly different marginal costs choose what quantity to make of a homogeneous product whose price is determined by the intersection of the product demand curve and the joint production of the firms. This model implies a positive relationship between market concentration and average market power. More concentration implies less competition. In effect, with fewer firms, each firm has less competition to take into account and more ability to raise price above marginal cost. Furthermore, in this model, welfare is lower in more concentrated markets because of the deadweight loss associated with price-cost margins.[3]

On the other hand, a large class of commonly used industry models predict a positive relationship between competition and concentration. All involve heterogeneous-cost firms selling differentiated goods. The models build in this differentiation in various ways, ranging from a direct preference parameter to trade, transport, or search costs. More substitutability implies firms' residual demand curves are more elastic. Reflecting this heightened competition, price-cost margins are lower when substitutability is high. Examples of models in this class include Melitz (2003), Asplund and Nocke (2006), Melitz and Ottaviano (2008), and Foster, Haltiwanger, and Syverson (2008).[4]

[3] Specifically, under this framework, one can show that the share-weighted sum of firms' Lerner indexes—their price-minus-marginal-cost margin as a share of the price—equals the Herfindahl–Hirschman index divided by the price elasticity of demand. In this issue, Berry, Gaynor, and Scott Morton offer additional discussion of interpreting the connections between competition and the Herfindahl–Hirschman index in the Cournot model.

[4] Versions of these types of models where the differentiation is instead in product quality are often isomorphic to the heterogeneous-cost version.

In such models, actual market entrants come from a pool of potential entrants who decide whether to pay a sunk entry cost to draw a cost level from a known distribution. Entrants who choose to receive a draw determine after observing the draw whether to begin production and thus to earn the corresponding operating profits. This setup creates a threshold cost level where only potential entrants receiving sufficiently low-cost draws enter in equilibrium. This basic structure implies a comparative static result where increases in substitutability (consumers are more willing or able to shift to different producers) both reduce margins and make it harder for higher-cost firms to operate. Additionally, because consumers are more responsive to any given cost difference, the responsiveness of market shares to cost differences is larger when substitutability rises. Thus, an increase in substitutability both *reduces* price-cost margins and *increases* concentration. In contrast to the Cournot case, the model predicts a negative correlation between market power and concentration.

Two other predictions of these models are relevant to this discussion. First, welfare rises along with substitutability; heightened competition reduces margins and the associated deadweight loss. Second, as substitutability/competition increases, profits of the firms operating in the market actually *increase*. In the theoretical model, more intense competition reduces the range of operating cost draws that are profitable, reducing successful entry rates. As a result, profits conditional on operating must rise to counterbalance the higher risk of failure. Interestingly, then, higher profits among firms in the market are a sign not of less competition, but more. Profits rise, despite the lower margins, because quantities sold increase markedly as substitutability/competition rises. Models in this class emphasize the earlier point that concentration is an outcome of underlying forces of supply and demand that can play out in various ways.

A negative relationship between market power and concentration is not just a theoretical curiosity. Many empirical studies in varied settings have found that greater substitutability/competition—resulting from, say, reductions in trade, transport, or search costs—shifts activity away from smaller, higher-cost producers and toward larger, lower-cost producers. As an example, in Syverson (2004a, b), I show that increases in the ease with which consumers can substitute among producers—spatial differentiation is limited, or products are more physically similar—force out the least efficient producers and increase skewness in the size distribution. In Goldmanis et al. (2010), we demonstrate that search cost reductions reallocate market share toward lower-cost and larger sellers, increasing market concentration even as margins fall. It is not an exaggeration to say that there are scores, perhaps hundreds, of such studies. Some focus on specific industries; others are broader.[5] Perhaps most

[5] Changes in production technologies that increase scale economies can also raise concentration. Unlike increases in product substitutability, which by their nature tend to flatten residual demand curves and therefore reduce market power, scale economies have no direct influence on demand. Thus, their equilibrium effect on market power is more ambiguous. However, prices could very well still fall even if markups do not, because the scale economies have reduced marginal costs. Arguably, this mechanism in part accounts for the transformation of the US retail sector over the past several decades, first through

relevant to the current discussion are Autor et al. (2017) and Crouzet and Eberly (2019). These studies find patterns of simultaneous concentration and productivity growth in settings that speak very directly to the recent macro market power literature as well. I will discuss each in more detail below.

In thinking about concentration as a measure of market power, there is a sharp split between the macro and micro market power literatures. From the 1950s through the 1970s, industrial organization often tried to link measures of market concentration to the behavior of firms and to resulting profit, in what was known as the structure-conduct-performance literature. But by the 1980s, given the very real concerns that concentration was likely to be misleading as a measure of market power, the field of industrial organization essentially stopped comparing market outcomes such as prices, margins, and profit rates to concentration levels—especially when making comparisons across markets or industries that differ in demand and technology fundamentals. While I would not call for a blanket ban on the practice of using concentration to measure market power, caution about the practice is well warranted. There were good reasons for industrial organization to choose to forgo it (particularly, again, for cross-industry comparisons). Simply put, the relationship between concentration and markups, prices, or profits is a relationship between market outcomes. These can be uninformative or, worse, misleading about the causal effect of competition.

Below I will speak further to what the microeconomic literature typically does to measure market power, whether it is practical for macro-oriented work, and what other alternatives might be available.

Direct Measurement of Markups with Accounting Data

To estimate markups directly, one needs data on prices and marginal cost. Data on prices is relatively straightforward to obtain, but data on costs across a wide range of firms—and especially data on marginal costs—is harder to find. One approach here is to use what researchers refer to as "accounting data"—essentially, data from firms' financial statements—and then work with this data to develop estimates of marginal costs. In two recent papers, De Loecker and Eeckhout (2018) and De Loecker, Eeckhout, and Unger (2018) take this approach to estimating price–marginal cost markups in the United States and around the world. In the US-centric study, they use the Compustat database, comprising the harmonized financial reports of publicly listed companies for the past several decades (De Loecker, Eeckhout, and Unger 2018). The world study uses Thomson Reuters Worldscope, which spans over 100 countries and contains income statements mostly for publicly traded companies, though it does also include some private firms (De Loecker and Eeckhout 2018).

the diffusion of warehouse centers and superstores and more recently through e-commerce (Hortaçsu and Syverson 2015).

The simplest method with which one might use accounting data to measure markups is to look at the ratio of revenues to total variable costs, which—when both of these are divided by quantity produced—would be equal to the ratio of price to average variable cost. Average variable cost does not of course generally equal marginal cost, but marginal cost is very hard to measure directly. Only when marginal cost is constant at all quantity levels is it equal to average variable cost. Moreover, accounting cost categorizations do not make it easy to separate variable from fixed costs on a consistent basis.

Thus, the studies by De Loecker and Eeckhout (2018) and De Loecker, Eeckhout, and Unger (2018) move beyond the simple proxy approach to obtain a more sophisticated estimate of markups. They use a firm-level variant of a method Hall (1988, 2018) developed and applied to industry-level data.[6] Hall (1988) shows that under cost minimization, for any variable input (an input that is freely adjustable by firms within any given period, as opposed to an input that is quasi-fixed, as many forms of capital are often thought to be), the firm's markup will equal the ratio of two values: the elasticity of output to that variable input, and the share of revenues the input is paid. That is,

$$\mu = \frac{\beta_v}{s_v},$$

where μ is the (multiplicative) markup, β_v is the elasticity of output with respect to the variable input v (from the firm's production function), and s_v is the share of revenues paid to the variable input supplier. Basu (in this issue) overviews this and related "production-function-based" approaches for measuring markups.

The accounting data include a measure called "cost of goods sold" (COGS). De Loecker, Eeckhout, and Unger (2018) use this as a measure of variable inputs. They estimate a production function by regressing revenues on this measure of COGS and on the book value of capital for all firms in an industry. This yields an estimate of the elasticity of output with respect to COGS.[7] The other piece of information necessary to estimate the markup, the share of revenue paid to this category of COGS, is observed directly in the data. They take the quotient of these two elements to obtain an estimate of the markup for every firm-year in their data. (The elasticity β_v is restricted to be the same across all firms in an industry or industry-year depending on the specification. The revenue share s_v is firm-year specific.)

[6] Hall (2018) uses industry data and finds mixed support for increasing markups. He estimates an average trend in measured markups between 1988 and 2015 that is positive but statistically insignificant (0.6 percent annual growth, with a standard error of 0.5 percent). Multiple measures of returns to capital rise. There is a low correlation between the levels and growth rates of three measures of market power he constructs, but there is a modest positive correlation between concentration and measured markups in his sample.

[7] Production function estimation is itself the subject of a large methodological literature and raises additional measurement issues beyond the scope of our discussion here.

An attention-getting headline number from De Loecker, Eeckhout, and Unger (2018) is that the revenue-weighted average markup in the United States climbed from about 1.2 in 1980 to 1.6 in 2014. They also find increasing skewness in the across-firm distribution of markups over that period, with average markup growth coming from a spreading of the right tail and a shift in revenue shares toward higher-markup firms. Indeed, the median firm-level markup remained essentially constant throughout the time period.

In their study with international data, De Loecker and Eeckhout (2018) find a similarly sized increase in the size-weighted markup, from 1.1 in 1980 to 1.6 in 2016. Some systematic variations in this trend exist across continents, however. While Europe, North America, Asia, and Oceania saw rather steady increases over 1980–2015, average markups in South America had little discernible trend. Markups in Africa jumped up between 2000 and 2005 but were level before and after.

One of the most compelling elements of these studies is that they are using a direct measure of price-cost margins to gauge market power. In terms of vulnerabilities, accounting data are not constructed for the sake of measuring economic categories like variable costs. Accounting data include two primary categories of costs: (1) cost of goods sold and (2) selling, general, and administrative (SG&A) expenses. COGS includes direct costs associated with purchasing and transforming inputs into the product a company sells and as such is thought to be composed primarily of variable costs. The SG&A category includes most other costs and as such captures many fixed costs. That said, some SG&A expenses might plausibly scale with the size of operations, while some costs in COGS might arguably be fixed. Indeed, accounting standards actually allow classification of expenses by COGS and SG&A to vary by sector. In the end, the variable/fixed demarcation is not as clean as one would like it to be for measuring markups.

How one measures variable costs matters empirically. Traina (2018) shows that if the sum of COGS and SG&A is used as the variable input measure, both the estimated levels and, more to the point, the changes in US markups fall. Instead of rising from 1.2 to 1.6 over 1980–2015, Traina's alternate markups rise from only around 1 to 1.15. Of course, this estimate of markup growth could itself be flawed because of the imperfect mapping between accounting and economic cost categories, as De Loecker, Eeckhout, and Unger (2018) point out. In the end, researchers using this approach are left to make choices among imperfect options.

A separate measurement (and conceptual) issue is that while this ratio of marginal product to revenue share equals the output markup under the assumptions of imperfect competition in the product market and a perfectly competitive market for the variable input, it also equals the monopsony *markdown* in the wage of the variable factor if instead the product market is perfectly competitive and producers have market power in the factor market. If firms have market power in both the product and factor markets, then the ratio reflects the combination of these two effects. Therefore, reading the ratio as reflecting solely product market monopoly (or only factor market monopsony, for that matter) could misattribute one for the other. Moreover, even recognizing that the measured ratio may reflect both market power effects, separately quantifying each component requires additional variation and empirical metrics beyond simply constructing the ratio.

Posing a Paradox: Markups and Their Relations

A widespread change in markups across an economy will necessarily have implications for other macroeconomic variables, including price inflation, cost growth, and profits. In Syverson (2018), I raise a potential inconsistency among measures of inflation, markups, and cost growth. I can summarize the paradox using the relationship that price, P, equals a markup rate, μ, times cost, C:

$$p = \mu \cdot C.$$

According to firms' profit-maximization theory, the relevant cost C ought to equal marginal cost, and the markup μ should be a function of consumers' price sensitivity. However, even if prices are not set to maximize profits, the relationship is still quite general and useful. For any consistently measured price and cost, one can define the markup μ as whatever multiplicative factor makes the relationship hold (μ could even be less than 1, if price is less than cost for some reason). In this sense, the relationship is essentially an identity.

The same relationship applies to growth rates. That is,

$$\text{Growth in } P \approx \text{Growth in } \mu + \text{Growth in } C.$$

Expressing the relationship in growth rates is an approximation, but it will be close to exact in the situations in which we are interested, where growth rates are relatively modest.

Now consider the empirical patterns observed in each of these growth rates over the past few decades. The left-hand side, the growth rate of prices, is inflation. Measured inflation has been low over the past few decades, especially relative to what many consider as its traditional driving forces. The first term on the right-hand side, the growth rate of markups, has been estimated to be quite high in some studies, which is the object of focus here. But if price growth is relatively low and markups are growing quickly, costs must be falling quickly. It is not clear in the data that this is the case.

Two factors affect the growth rate of costs: productivity and factor prices. Productivity growth has been in a slump since the mid-2000s. Productivity is inversely related to costs, so when productivity grows more slowly than usual, cost growth will tend to be higher than usual. As for factor price trends over the past couple of decades, wage growth has been slow, if anything (more so for the middle and lower end of the distribution than for the high end), and interest rates have fallen to historic lows. In isolation, those factor price patterns would tend to slow the growth rate of cost, but they are countervailed by slowing productivity growth.

We can investigate the net effect of these two influences by looking at "unit costs," which conveniently combine both productivity and factor price effects on costs. Unit labor costs are the ratio of total compensation per hour worked to labor productivity—the nominal labor compensation required to build one unit of output. According to Bureau of Labor Statistics data, the growth in US aggregate unit labor cost has been somewhat slower than inflation (this is also reflected in the labor compensation's falling share of income). This opens the possibility that low labor cost growth has "made room" for higher markups. However, the timings of the two trajectories do not line up well. Much of the decline of labor's share has occurred since 2000, while the period with the fastest increases in markups was 1980–2000. Furthermore, nominal unit cost growth for other factors may have accelerated over the period. The average growth rate of the "unit nonlabor payments" series from the Bureau of Labor Statistics, which includes capital payments, taxes on production, and profits, has slowly risen during the past two decades. This could in part reflect profits from increased markups showing up in addition to actual costs. Unfortunately, the Bureau of Labor Statistics does not break profits out of unit nonlabor payments for the entire nonfarm business sector. They do for nonfinancial corporations, however. These indicate increasing unit capital costs for such firms, with unit nonlabor costs having grown faster than inflation for the past 20 years.

In short, measures of costs do not seem to behave in the way implied by the measured trends in inflation and markups. One potential resolution of the paradox comes from parsing types of cost. Productivity and unit cost measures probably most closely reflect average cost. If marginal costs were rising at a slower rate than average costs, it is possible that unit cost growth could be steady even as inflation remained unusually low. The former would reflect steady changes in average cost; the latter would reflect faster reductions in marginal cost.

This story has the right qualitative features to resolve the paradox. However, it is unclear that it can quantitatively account for the differential patterns in prices, markups, and costs. A decomposition of the price-cost markup, first made by Susanto Basu in an earlier discussion of De Loecker and Eeckhout's work, is instructive about this.[8]

Rewrite the markup expression by multiplying and dividing it by average costs:

$$\mu \equiv \frac{P}{MC} = \frac{P}{AC}\frac{AC}{MC}.$$

Multiplying and dividing *P/AC* by the output quantity makes it clear that the markup is equivalent to the ratio of revenues to total costs. The *AC/MC* ratio is, by definition, the scale elasticity of the function that relates a firm's costs to its output (that is, the inverse of the elasticity of costs with respect to quantity).[9] When marginal costs are

[8] I am grateful to Susanto Basu for conversations regarding this decomposition.

[9] This is straightforward to verify. The elasticity of costs with respect to quantity for any differentiable cost function $C(Q)$ is $C'(Q)(Q/C) = MC(1/AC)$. For homothetic production functions, the scale elasticity equals the returns to scale of the production function. Note that $C(Q)$ is the function that relates the

less than average costs, average costs are falling in quantity and the scale elasticity is greater than one. Conversely, if $MC > AC$, there are diseconomies of scale, and the scale elasticity is less than one.

We therefore have, using ν to denote the scale elasticity,

$$\mu = \frac{R}{TC}\nu.$$

Define pure profit's share of revenues as

$$s_\pi \equiv \frac{R - TC}{R}.$$

We can rewrite the markup as

$$\mu = \frac{1}{1 - s_\pi}\nu.$$

Thus, the markup must equal the inverse of one minus profit's share of revenue times the scale elasticity. Note that the only assumption required to derive this expression is that the cost function $C(Q)$ is differentiable; the other manipulations were just algebra or identities.

This expression reveals an empirical discipline on measures of markups at the firm level. Namely, markup levels must also imply something about profit shares, scale elasticities, or both. If a firm sees a substantial increase in markups over time, there must also be an increase in pure profit's share of its income or in its scale elasticity.

It is often difficult to obtain firm-level estimates of scale elasticities, as common technologies are typically imposed across firms by researchers in order to estimate an elasticity. Thus, investigating the markup profit-share scale-elasticity relationship firm by firm can be hard. Exploring its aggregate version can still be informative, though this should be accompanied by the caveat that, as they are ratios, the nonlinearity of the markup and the scale elasticities implies that weighted averages of firm-level values will not generally exactly add up to their aggregate analogs. I make this aggregate comparison here, noting this proviso.

As noted above, DeLoecker, Eeckhout, and Unger (2018) report that US average markups grew from 1.21 to 1.61 between 1980 and 2016.[10] Suppose that the production technology remained stable enough over the period so that the scale

firm's production cost to its output in practice; it need not necessarily be *the* cost function (that is, the one that assumes the firm has chosen the minimum-cost bundle of inputs required to make any given Q). Thus, the expression still applies even if firms aren't cost-minimizing.

[10] De Loecker, Eeckhout, and Unger (2018) report several sets of markup estimates. I am using their benchmark "PF1" specification. The alternative estimates exhibit quantitative behaviors similar to those described here.

elasticity didn't change. Then pure profit's share of revenues in 2016 must be the following function of its 1980 share:

$$\frac{1.61}{1.21} = \frac{\dfrac{1}{1 - s_{\pi,2016}}}{\dfrac{1}{1 - s_{\pi,1980}}},$$

$$s_{\pi,2016} = 0.25 + 0.75\ s_{\pi,1980}.$$

Even if profit's revenue share was zero in 1980, the observed change in markups—in the absence of any increase in scale economies—would imply the profit share in 2016 would be 25 percent. Given that this is a share of revenue and that total aggregate revenues (that is, sales) are roughly double aggregate value added, this implies that profits were roughly half of all value added in 2016. This would be unrealistically large. By any measure, labor's share of value added is greater than this. Even if we were to consider *all* capital income as pure economic profit (that is, capital's competitive return was zero), the increase in measured markups does not make empirical sense in the absence of substantial changes in scale economies.

What if scale economies did increase over the period? Fixed costs may have grown, or the output product mix may have shifted in composition toward products with lower marginal costs (like software and pharmaceuticals). An empirical test is feasible here. Pure profit rates can be estimated, although doing so requires assumptions about how to measure capital's competitive return. One can estimate production functions to obtain scale elasticities. The recent literature contains some estimates along these lines. Barkai (2017) constructs a measure of pure profit's share of value added, finding that from 1908 to 2014 it grew from 3 to 16 percent, and its revenue share thus grew from about 1.5 to 8 percent. (I will discuss this study in detail below.) De Loecker, Eeckhout, and Unger (2018) estimate changes in the average scale elasticity for firms in their sample, finding that it rose from 1.03 to 1.08 during 1980–2016. Plugging these values into the relationship above and taking their ratio yields:

$$\frac{\mu_{2016}}{\mu_{1980}} = \left(\frac{1 - s_{\pi,1980}}{1 - s_{\pi,2016}}\right)\frac{\nu_{2016}}{\nu_{1980}},$$

$$\frac{1.61}{1.21} = \left(\frac{1 - 0.08}{1 - 0.015}\right)\frac{1.08}{1.03},$$

$$1.33 = (1.08)1.05,$$

$$1.33 = 1.14.$$

While the relationship is still some distance from implying consistency, it is closer to equality, suggesting that growth in scale economies is part of the story. In addition, there is the caveat that I am mixing aggregates and firm-level averages when the relationship should hold firm by firm.

The relationship between markup, profit share, and scale elasticity is a tool that can be applied more generally, whether among firms in the cross section or over time. While there are practical hurdles, estimates of the necessary components are generally feasible to obtain in the data. The relationship imposes a useful consistency check on empirical estimates in this area.

Market Power and Low Investment Rates

Corporate profit rates and Tobin's *q* (the ratio of a firm's market value to the book value of its assets) have both been relatively high since 2000. However, during the same period, the investment rate has been low relative to its historical connections to profits and Tobin's *q*. In a pair of papers, Gutiérrez and Philippon (2017a, b) marshal evidence suggesting market power may be behind the low investment rates.

Gutiérrez and Philippon (2017b) run a horse race between alternative hypotheses for low investment: a rise in financial frictions, changes in the nature of investment (like intangibles replacing measured capital investment or globalization shifting investment abroad), increased short-termism in management, and decreased competition. Each class of explanation has multiple specific measures. They find that, at least in terms of ability to explain statistically the unusually low observed investment rate, the rising importance of intangibles accounts for about one-third of the drop, while corporate ownership structure (what fraction of company stock is held by likely long-term investors) and increased industry concentration explain the rest. In their framework, measures of financial frictions have no explanatory power.

To address the question of causality more directly, Gutiérrez and Philippon (2017a) use natural experiments and instrumental variables techniques to link changes in competition to investment. The natural experiments involve two measures of increased competition from Chinese imports. The instrumental variable is a measure of "excess entry" in an industry in the 1990s. The logic of the instrument is that the go-go US startup environment of the latter part of that decade in particular led to a large amount of essentially random volatility in entry rates across markets. They show that the amount of 1990s entry relative to fundamentals (both current and in expectation) is correlated with industry concentration a decade later, but uncorrelated with observable shocks that occurred in the interim. Instrumenting for industry concentration using excess entry, they find that concentration is negatively correlated with investment rates.

These papers are persuasive in their case that measured investment is low relative to standard explanatory variables. Moreover, they demonstrate the *potential* for market power not only to create inefficiencies and reduce output today but also, through its investment effects, to reduce future growth rates. But I believe the case is not yet proven. A few critiques present themselves here.

If intangible capital has become more important over the past couple of decades, and the composition of investment has shifted toward it as a result, the quantitative response of *measured* investment (which of course does not include intangibles) to traditional variables like corporate profits and Tobin's *q* would decline. But this would be a measurement change, not necessarily an economic one. Gutiérrez and Philippon (2017b) do consider this possibility, and they seek to address it using proxies for intangible capital (including "tangible intangibles"—the capitalized R&D, software, and artistic originals series constructed by the Bureau of Economic Analysis). Their proxy for intangibles does explain some of the drop in the measured investment rate, but given the uncertainties involved in measuring intangible investment, a different proxy might explain still more.

However, a deeper issue is that intangible investment need not just be associated with (or caused by) concentration; in addition, it can causally affect industry concentration. Thus, intangibles aren't just another factor in addition to concentration that might explain low measured investment. They might be affecting concentration directly.

Crouzet and Eberly (2019) point out that an intangibles-concentration connection can occur through two mechanisms, with very different economic implications. One is that increased concentration, in this case reflecting less competition, reduces the incentives of firms to invest, and this might be coincidentally (or perhaps even causally) correlated with growth in intangible intensity. The other mechanism reverses the potential causality between intangibles and concentration and has diametrically opposed implications for welfare. If a company invests in intangibles that allow it to deliver a higher-quality product at a lower price (by reconfiguring its organizational structure and internal processes, for example), market share will naturally shift toward that company, creating coincident growth of intangible intensity and industry concentration. However, this rise in concentration would be efficiency enhancing, as the total resources required to deliver a given amount of product quality (and consumer welfare) would have fallen.

What suggestive evidence might be brought to bear on these two possibilities? Intangible investment intensity is highest and grows fastest for the largest and fastest-growing firms in an industry, according to the estimates from Crouzet and Eberly (2019). In addition, they compare sector-level trends in labor productivity levels and Hall's (2018) industry-level markup estimates. They find that within the manufacturing and consumer sectors (the latter combining wholesale and retail trade as well as agriculture), estimated markups were flat over 1990–2015, while labor productivity rose. They therefore attribute the coincident increases in concentration and intangible intensity observed in those sectors to efficiency-enhancing mechanisms. On the other hand, they find that both markups and labor productivity grew steadily in the health-care and high-tech sectors, indicating elements of both market power and efficiency gains at work.

These results suggest that the connections between lower measured investment and concentration do not reflect an across-the-board influence of rising market power but instead are an amalgamation of differing mechanisms with quite different economic interpretations. Sector-specific mechanisms would be consistent with, for example, the notions that globalization has increased competitive pressures in manufacturing (Feenstra and Weinstein 2017) while merger waves have reduced competition in healthcare (Gaynor 2018). Some recognition of the heterogeneity underlying aggregate patterns seems clearly warranted, both for understanding the phenomenon and for drawing welfare implications. This insight also vividly evokes the aforementioned issues involved with assuming that concentration is a useful measure of market power.

Market Power and the Labor Share of Income

It seems fair to say there is a consensus in the profession that labor's share has been trending down, while corporate profits have risen. While some have raised concerns about specifics of measurement, the trends have been documented in multiple ways. The macro market power literature has raised the possibility that these changes may be related to a rise in market power, markups, and pure profit.

In an example of research along these lines, Barkai (2017) decomposes aggregate factor income into three elements: labor's share, capital's share, and pure profit. Labor income is taken directly from national income accounts and is therefore measured in standard ways. To compute capital income, Barkai multiplies the observed aggregate capital stock by a user cost of capital. The user cost equals a real interest rate (constructed as average blue-chip bond yields in a period minus a measure of expected inflation) plus a measure of the depreciation rates. This user cost is supposed to reflect the competitive return earned by capital inputs. Any remaining income is considered pure profit. The results of this approach indicate that the drop in the share of income paid to labor was accompanied by a slight drop in capital's share. Meanwhile, the pure profit residual increased substantially, from 3 percent of national income in 1985 to 16 percent by 2014.

The study ties this shift in factor income shares to market power, using regressions conducted at the six-digit North American Industry Classification System (NAICS) level. Industries that saw larger increases in concentration saw bigger drops in labor's share of income. Barkai (2017) interprets this as evidence that declining competition has been responsible at least in part for the secular decline in labor's share.

Eggertsson, Robbins, and Getz Wold (2018) also emphasize the pure profit approach by augmenting a standard neoclassical model with increasing market power/markups. They show that, suitably parameterized and in the presence of a decreasing natural rate of interest, the model can qualitatively and quantitatively explain the falling labor's share of income as well as several other phenomena: the increase in the pure profit rate, growth in the financial wealth-to-output ratio, an increase in Tobin's *q* without associated investment, and a divergence between the marginal and the average return on capital. The mechanism, briefly stated, is that growing market power (what the paper terms the "emergence of a non-zero-rent economy") leads directly to the increase in pure profit through higher markups. Financial wealth and Tobin's *q* both reflect future claims on profits, so these rise as well. The increase in pure profit's share decreases both labor's and capital's shares.

Because higher profits increase the return on capital, however, there must be a countervailing influence in order to generate the roughly constant returns that have been observed in the data. This is where the falling natural rate of interest comes in.

Other papers start with the same patterns but emphasize different explanations. The model of Farhi and Gourio (forthcoming) shares several basic structural elements with Eggertsson, Robbins, and Getz Wold (2018), but the analysis allows for and emphasizes the role of a changing risk premium in explaining several of the observed patterns. Relatedly, Karabarbounis and Neiman (2018) argue that a shrinking labor share and rising pure profit (what they call "factorless income") can best be explained by a rising rental rate for capital. They point out that explaining a lower labor share of income via higher pure profit and/or returns to intangible capital, while consistent with many empirical patterns in recent decades, implies implausible empirical patterns when applied to the 1960s and 1970s. The mismeasured rental rate explanation avoids these counterfactual predictions and implies some other empirical patterns more consistent with observed data. At the same time, the study lacks a sharp explanation for what would cause the true rental rate to vary as it would need to in order to produce the macroeconomic outcomes that have occurred over the prior 60 to 70 years.

Other papers raise the general point that concentration can be related to macroeconomic outcomes through mechanisms other than market power; in particular, in Autor et al. (2017) and Bessen (2017), higher concentration and lower labor income may be part of an efficiency-enhancing shift. Both papers argue that concentration has grown because changes in market factors have created an environment that increases skewness—in revenues (as measured by rising concentration) and productivity, certainly, and perhaps in other dimensions. Something has flattened firms' residual demand curves or marginal cost curves, be it increased scale economies, network effects, or improved abilities of consumers to find low-cost or high-quality firms. These changes lead to increased concentration ("superstar firms" in the parlance of Autor et al. 2017) but do not necessarily imply growth in market power. Increased scale economies may come from reductions in marginal cost that reduce the amount of inputs necessary to produce output—an efficiency enhancement. On the other hand, scale economies also require enough market power in equilibrium for firms to pay fixed costs and production costs of their inframarginal units. Network effects also have implications for both efficiency and market power. Consumers can obtain a utility benefit from network effects, but network effects can also cause lock-in, which gives firms pricing power. Improving consumers' abilities to choose from whom they buy—which may come from changes in search, transport, or trade costs, for example—is likely to be efficiency enhancing.

Both Autor et al. (2017) and Bessen (2017) present evidence bolstering the case for an efficiency-enhancing mode of concentration being the primary actor in their data. Autor et al. find that industries that saw greater increases in concentration also saw on average faster growth in patent rates, capital intensity, and productivity. Bessen ties use of information technology systems to concentration as well as to more skewed operating margins and productivity levels in an industry. To the extent that gains in concentration have been accompanied by efficiency gains, caution is again warranted when using concentration as a metric to infer market power.

Filling in the Macro Market Power Literature

In my discussion of various aspects of the macro market power literature, I have described some current vulnerabilities that in my opinion keep the literature's conclusions from being dispositive. What might be done to fill holes and round out the evidence in a way that would allow more definitive conclusions?

One logical place to look for new threads that the literature could pick up is in the best practices of the well-developed microeconomic literature on market power. This literature typically starts with a recognition that the optimal price-cost markup depends on the slope of the inverse residual demand curve facing the firm. If that slope can be estimated, the implied profit-maximizing price–marginal cost margin can be backed out from that. Most of the microeconomic literature follows this logic and estimates the demand system for the products in the market (if the products are differentiated, this is typically accommodated by using a discrete choice demand system where the product attributes are included as demand shifters). While using demand-side data to infer marginal costs may seem surprising, in many settings the richness of the demand system offers the ability to estimate demand with some precision, and therefore implied margins, in ways that cost data alone could not. Moreover, one can typically jointly estimate both the demand and supply sides by parameterizing costs (again as a function of attributes if products are differentiated) and by using the restriction that the observed product price must equal the estimated marginal cost times the profit-maximizing markup implied by estimated residual demand.

However, the demand-system-estimation approach may not be feasible in the macro market power literature, with its broad combinations of industries and market settings. Specifying a realistic demand system typically takes a fair amount of knowledge about the nature of the product and the institutional details of the market. It is not practical to do this in studies that look across hundreds of very different markets. Taking an analogous approach with aggregate data would not work either. Pricing power depends on the slope of firms' residual demand curves, not the slope of the market/industry demand curve. Backing out an implied markup and market power from a market/industry demand curve would be conceptually and empirically incorrect.

If microdata are available, one might imagine a more parametric approach to measuring marginal cost whereby a cost function is specified and estimated using observed variation in costs. Marginal costs would then be derived from this estimated function. This approach is limited by several factors, however. Many producer-level datasets report only revenues, which combine quantity and price, and quantity data of some type are required to estimate a production function.

For highly differentiated products, there may not be enough data to characterize the cost function fully, given the multiple attributes that could shift costs. Also, to estimate a cost curve, instruments that exogenously shift quantities are needed, and these are not easy to find in many settings.

What is one to do, then? When it comes to estimating markups or measures of market power for broad swathes of the economy, there may be no silver bullet. One is left with a menu of imperfect choices.

Sometimes one can obtain direct measures of plausibly exogenous differences in competition. In that case, concentration might be instrumented using those measures, or alternatively, those measures could be used directly as explanatory variables themselves.

If there does not seem to be an alternative to concentration as a measure of market power, researchers should strive to demonstrate using ancillary evidence that increases in concentration do in fact correspond to more market power rather than efficiency in the market(s) they are studying. As an example of how this distinction might be made, Autor et al. (2017) show that concentration is associated with innovation, capital deepening, and productivity, which bolsters the case for efficiency mechanisms. Alternate findings from this methodology would have supported a market power interpretation.

Another area for ongoing research in the macro market power literature is to characterize heterogeneity more fully, both across and within markets. As mentioned above, the results from Crouzet and Eberly (2019) suggest that market power can act broadly within some sectors but not others. They find the health-care sector, for example, seems to have seen the influence of market power, while the manufacturing and consumer sectors are not showing the signs of market power. In turn, broad analysis across sectors can be compared to market-specific studies in the micro literature; for example, Cabral, Geruso, and Mahoney (2018) and work described in Gaynor (2018) support a finding of rising market power in the health-care sector. Characterizing such sectoral differences and explaining where they come from is important for understanding the mechanisms behind, and the effects of, market power in macro settings.

Within industries, the skewness results shown by De Loecker, Eeckhout, and Unger (2018)—that the increase in average measured markups is driven exclusively by increases in the right tail of the distribution—are an example of the necessity of understanding within-industry heterogeneity. Averages can obscure. Producers in an industry differ markedly in their behavior, including in their responses to common external influences. Market-, industry-, or economy-wide changes do not always, nor likely even usually, reflect a common change experienced by all producers. Rather, they reflect the summation of what are typically very different responses, which includes reallocations of activity across heterogeneous producers. The experience of the median producer (or even the average producer, if producers are equally weighted) may not be informative about changes at the industry level. One cannot simply rely on producer-level variation "canceling out" when looking at aggregate changes. That variation *is what creates* the aggregate changes.

Conclusion

The macro market power literature has offered an immense service by documenting and emphasizing the potential connections between several trends: labor's declining share of income, increasing corporate profits, increasing margins, increasing concentration, slower productivity growth, decreasing firm entry and dynamism, and reduced investment rates. While none of these is a perfect metric for market power, many (but not all) have been replicated in multiple venues with multiple techniques and as such can be considered reasonably robust. The fact that these changes are so noticeable and have been trending for so long (each for over a decade at a minimum, some approaching four decades now)—often in contrast to very different patterns before—creates an inherent interest and importance.

The market power story is very much a viable candidate explanation for the documented trends, especially in specific industries or sectors. However, I believe more evidence is yet required to make a broad-based increase in average market power the undisputed leading candidate explanation. Empirical gaps still need to be closed. There are plausible alternative stories, some accompanied by controverting empirical evidence to the market power hypothesis, that need to be rejected. Ultimately, indeed, it may be that the sources of the patterns are multicausal—some combination of greater intangible intensity, changing product-market substitutability, greater scale economies, and higher entry costs, all with potential implications for market power (though in possibly different directions). Moreover, the relative contribution of each could vary across sectors. Regardless, stronger conclusions will be warranted if researchers can make further progress in both qualifying and quantifying the roles of market power and alternatives.

■ *I thank Susanto Basu, Steven Berry, Judy Chevalier, Dennis Carlton, Jan De Loecker, Jan Eeckhout, Gauti Eggertsson, Marty Gaynor, Germán Gutiérrez, Bob Hall, Adam Looney, Thomas Philippon, Fiona Scott Morton, Carl Shapiro, and the editorial staff of this journal for comments.*

References

Asplund, Marcus, and Volker Nocke. 2006. "Firm Turnover in Imperfectly Competitive Markets." *Review of Economic Studies* 73(2): 295–327.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2017. "The Fall of the Labor Share and the Rise of Superstar Firms." NBER Working Paper 23396.

Azar, José, Ioana Marinescu, and Marshall I. Steinbaum. 2017. "Labor Market Concentration." NBER Working Paper 24147.

Azar, José, Martin C. Schmalz, and Isabel Tecu. 2018. "Anticompetitive Effects of Common Ownership." *Journal of Finance* 73(4): 1513–65.

Barkai, Simcha. 2017. "Declining Labor and

Capital Shares." https://www.london.edu/faculty-and-research/academic-research/d/declining-labor-and-capital-shares.

Benmelech, Efraim, Nittai Bergman, and Hyunseob Kim. 2018. "Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?" NBER Working Paper 24307.

Bessen, James E. 2017. "Information Technology and Industry Concentration." Boston University School of Law, Law and Economics Research Paper 17-41. https://ssrn.com/abstract=3044730.

Cabral, Marika, Michael Geruso, and Neale Mahoney. 2018. "Do Larger Health Insurance Subsidies Benefit Patients or Producers? Evidence from Medicare Advantage." *American Economic Review* 108(8): 2048–87.

Crouzet, Nicolas, and Janice C. Eberly. 2019. "Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles." NBER Working Paper 25869.

De Loecker, Jan, and Jan Eeckhout. 2018. "Global Market Power." NBER Working Paper 24768.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2018. "The Rise of Market Power and the Macroeconomic Implications." https://sites.google.com/site/deloeckerjan/research/RMP-DLEU.pdf.

Eggertsson, Gauti B., Jacob A. Robbins, and Ella Getz Wold. 2018. "Kaldor and Piketty's Facts: The Rise of Monopoly Power in the United States." NBER Working Paper 24287.

Farhi, Emmanuel, and Francois Gourio. Forthcoming. "Accounting for Macro-Finance Trends: Market Power, Intangibles, and Risk Premia." *Brookings Papers on Economic Activity*.

Feenstra, Robert C., and David E. Weinstein. 2017. "Globalization, Markups, and US Welfare." *Journal of Political Economy* 125(4): 1040–74.

Foster, Lucia, John Haltiwanger, and Chad Syverson. 2008. "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review* 98(1): 394–425.

Furman, Jason, and Peter Orszag. 2015. "A Firm-Level Perspective on the Role of Rents in the Rise in Inequality." Presentation at "A Just Society" colloquium in honor of Joseph Stiglitz, Centennial Event, Columbia University, New York, NY, October 16–17, 2015.

Gaynor, Martin. 2018. "Examining the Impact of Health Care Consolidation." Statement before the Committee on Energy and Commerce Oversight and Investigations Subcommittee, US House of Representatives, February 14, 2018. https://docs.house.gov/meetings/IF/IF02/20180214/106855/HHRG-115-IF02-Wstate-GaynorM-20180214.pdf.

Goldmanis, Maris, Ali Hortaçsu, Chad Syverson, and Önsel Emre. 2010. "E-commerce and the Market Structure of Retail Industries." *Economic Journal* 120(545): 651–82.

Goolsbee, Austan, Steven Levitt, and Chad Syverson. 2016. *Microeconomics*, 2nd ed. New York: Worth Publishers.

Gutiérrez, Germán, and Thomas Philippon. 2017a. "Declining Competition and Investment in the U.S." NBER Working Paper 23583.

Gutiérrez, Germán, and Thomas Philippon. 2017b. "Investment-less Growth: An Empirical Investigation." *Brookings Papers on Economic Activity* 2017(Fall): 89–182.

Hall, Robert E. 1988. "The Relation between Price and Marginal Cost in U.S. Industry." *Journal of Political Economy* 96(5): 921–47.

Hall, Robert E. 2018. "New Evidence on the Markup of Prices over Marginal Costs and the Role of Mega-Firms in the US Economy." NBER Working Paper 24574.

Hortaçsu, Ali, and Chad Syverson. 2015. "The Ongoing Evolution of US Retail: A Format Tug-of-War." *Journal of Economic Perspectives* 29(4): 89–112.

Karabarbounis, Loukas, and Brent Neiman. 2018. "Accounting for Factorless Income." NBER Working Paper 24404.

Khan, Lina M. 2017. "Amazon's Antitrust Paradox." *Yale Law Journal* 126(3): 710–805.

Krueger, Alan B. 2018. "Reflections on Dwindling Worker Bargaining Power and Monetary Policy." Luncheon Address at the Jackson Hole Economic Policy Symposium, Jackson Hole, WY, August 24, 2018.

Manning, Alan. 2003. *Monopsony in Motion: Imperfect Competition in Labor Markets*. Princeton, NJ: Princeton University Press.

Melitz, Marc J. 2003. "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity." *Econometrica* 71(6): 1695–725.

Melitz, Marc J., and Giancarlo I. P. Ottaviano. 2008. "Market Size, Trade, and Productivity." *Review of Economic Studies* 75(1): 295–316.

Pindyck, Robert S., and Daniel L. Rubinfeld. 2012. *Microeconomics*, 8th ed. Boston: Pearson.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter. 2018. "Diverging Trends in National and Local Concentration." NBER Working Paper 25066.

Scott Morton, Fiona, and Herbert J. Hovenkamp. 2018. "Horizontal Shareholding and Antitrust Policy." *Yale Law Journal* 127(7): 2026–47.

Shapiro, Carl. 2018. "Antitrust in a Time of

Populism." *International Journal of Industrial Organization* 61(1): 714–48.

Syverson, Chad. 2004a. "Market Structure and Productivity: A Concrete Example." *Journal of Political Economy* 112(6): 1181–222.

Syverson, Chad. 2004b. "Product Substitutability and Productivity Dispersion." *Review of Economics and Statistics* 86(2): 534–50.

Syverson, Chad. 2018. "Changing Market Structure and Implications for Monetary Policy." Remarks at the Jackson Hole Economic Policy Symposium, Jackson Hole, WY, August 24, 2018.

Traina, James. 2018. "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements." Stigler Center New Working Paper Series 17.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 44–68

Do Increasing Markups Matter? Lessons from Empirical Industrial Organization

Steven Berry, Martin Gaynor, and Fiona Scott Morton

Many economists and policymakers are expressing concern over the possibility of increasing monopoly power in the US and the world economy. There have been decades of research in industrial organization devoted to understanding how one can (and cannot) reliably learn about the causes and consequences of market power and markups—that is, a positive difference between price and marginal cost.

Starting about 30 years ago (Bresnahan 1989), the field of industrial organization adopted methods for understanding firm conduct and markets on the basis of the relevant economic primitives: demand, cost, and pricing conduct. Thus, under the assumptions that firms maximize profits and have to cover their total costs, the equilibrium price (and other outcomes, such as product choice, location, quality, and innovation) will be determined by demand, marginal costs, and fixed (possibly sunk) costs, along with the conditions of competition that shape pricing behavior. These conditions are modeled using modern game theory to incorporate imperfect competition, product differentiation, multiproduct firms, and firm entry, as well as a host of industry-specific institutions.

■ *Steven Berry is the David Swensen Professor of Economics, Yale University, New Haven, Connecticut. Martin Gaynor is the E. J. Barone University Professor of Economics and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania. Fiona Scott Morton is the Theodore Nierenberg Professor of Economics, Yale School of Management, New Haven, Connecticut. All three authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are steven.berry@yale.edu, mgaynor@cmu.edu, and fiona.scottmorton@yale.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.44 doi=10.1257/jep.33.3.44

However, a number of recent studies of markups instead employ an analytical approach that was broadly rejected by the field of industrial organization more than 30 years ago: the "structure-conduct-performance" paradigm. We begin by discussing the shortcomings of this approach, which involves regressions with an outcome such as markups or profits on the left-hand side and a measure of market concentration on the right-hand side, along with various control variables. This approach faces severe measurement problems and worse conceptual problems. As we will explain, there are numerous, quite different economic scenarios, with different welfare implications, that can result in a positive correlation between industry concentration and markups.

We then turn to some research that avoids the problems of the structure-conduct-performance approach. Although we mention several approaches, our main focus is on recent studies taking an industrial organization approach. As we will show, studies built on economic primitives sometimes describe a situation in which large firms are changing products and production methods, including the mix of marginal and fixed costs, over time. In some cases, the welfare effects for consumers are ambiguous; in others, larger firms seem to raise markups without a corresponding consumer benefit. In some of these cases, mergers may be playing a role in increasing markups. The strength of these industry-level studies is that they offer detailed insights into causes of higher markups; the corresponding downside is that without a surge of additional studies, it can be difficult to draw inferences about overall levels and trends in markups across the economy.

Building on these industrial organization studies, we summarize some of the main possible causes of expanding markups rooted in the underlying economic primitives. Possibilities include a rise in fixed or sunk costs, network effects, monopsony effects in labor markets, an increase in rent-seeking behavior, and globalization effects. As an example, higher fixed (or sunk) costs can lead to fewer firms in a market, which can result in softer competition, higher prices, and reduced consumer welfare. On the other hand, in some cases, higher fixed (or sunk) costs can be the endogenous outcome of improved products or of improved production technology that lowers marginal cost (Sutton 1991). In this case, observed higher markups may or may not be associated with higher prices and reduced consumer welfare.

In the final section of the article, we turn to antitrust enforcement and competition policy. We examine this not only because weakened antitrust policy offers a potential explanation for rising markups, but also because even if the main explanations lie elsewhere, antitrust policy offers some policy levers to address the rise in markups. Given the uncertainties about whether there has been an increase in markups and about the size of that increase, and given that these seem likely to vary across industries, our policy recommendations are focused on those that are beneficial under a wide range of conditions—for example, assuring that market entry is not blocked, that dominant incumbent firms don't engage in conduct to disadvantage rivals and harm competition, and that anticompetitive conduct in labor markets is not permitted. We offer the important caveat that regulatory, trade, and tax policies may also prove important in addressing any harms associated with increased markups.

Problems with Some Recent Studies of Market Power

Early empirical research in industrial organization from the 1950s into the 1970s employed the structure-conduct-performance paradigm to study how the extent of competition affected market outcomes. This empirical implementation of the paradigm typically involved regression analysis. The dependent variable was a market outcome such as profits, markups, or prices. The key explanatory variable sought to capture the structure of the market with a measure of concentration—usually the Herfindahl–Hirschman index, which is the sum of squared market shares. The regression also included a range of control variables intended to capture other exogenous reasons for variation. Structure is thus related to performance, with (unobservable) conduct captured as the estimated relationship between structure and performance. In this regression, the coefficient on the concentration measure is intended to capture how the toughness of competition changes as market concentration changes.

Within the field of industrial organization, the structure-conduct-performance approach has been discredited for a long time (Bresnahan 1989; Schmalensee 1989). But outside of industrial organization, the paradigm seems to have been readopted in recent years. Much of the recent attention to increasing markups or other market outcomes focuses on exactly this kind of evidence (for example, Furman 2015; Azar, Marinescu, and Steinbaum 2017; Barkai 2017; Bessen 2017; Gutiérrez and Philippon 2017a, b; Smith 2017; Azar et al. 2018; Benmelech, Bergman, and Kim 2018; Furman and Orszag 2018; Grullon, Larkin, and Michaely forthcoming). Such work sometimes proceeds without addressing the problems that led the field of industrial organization to reject the structure-conduct-performance approach.

Given the intuitive relationship between market concentration and firm performance, why did industrial organization reject the structure-conduct-performance paradigm? Researchers using the structure-conduct-performance approach were well aware of its limits at the time, as emphasized by Schmalensee (1989). We start with a discussion of measurement problems. The most important point, though, is that there are multiple causal paths that can explain a given correlation between concentration and other market outcomes. This implies that the very question—"What is the effect of concentration on prices or markups?"—is not well posed.

Measuring concentration is inherently difficult because economic markets are not observed directly in the data. For example, industrial classifications in the Census often fail to reflect well-defined economic markets. It is fairly clear that "software" is not a single industry, but much less clear how to divide it into separate industries. Other problems arise from geography. If Census data in an industry show a large number of small firms, this may represent a situation where they are in direct competition with one another or a situation where they operate in quite separate geographic or product markets. The Census does not measure degrees of product differentiation or homogeneity, or any measures of product-level prices.

Measuring economic outcomes was another problem for research in the structure-conduct-performance tradition. Most measures of profits use accounting measures, which are not economic profits. Markups are rarely directly observed in firm-level data at all, in part because firms' accounting structures are not set up to measure the economic concept of product-level marginal cost (Fisher and McGowan 1983). Attempts to estimate marginal cost involved additional, difficult measurement problems with regard to the size of fixed costs, sunk costs, and depreciation. In a best-case scenario, measured markups involve the markup of price over average variable cost.

Some researchers in the structure-conduct-performance tradition came to regressions using price as the dependent variable, rather than accounting profits or markups. But then, comparing prices across industries led to a call for industry-level structure-conduct-performance studies (Weiss 1990). Researchers understood that the nature of competition differs substantially from one industry to the next. For example, prices are determined in the food distribution industry via second price auction, in health care via bilateral bargaining, and in retail as posted prices. It's unclear what sorts of inferences are possible from estimates that aggregate across industries with such fundamental differences.

But even if the structure and output variables were measured with precision and the analysis was within a single industry, structure-conduct-performance researchers beginning with Demsetz (1973) often grappled with the problem of interpreting their regressions. For example, Ravenscraft (1983) regressed firm-level markups on firm market share and industry concentration, finding a coefficient on market share that was positive and significantly different from zero, but a near-zero (or even negative) coefficient on industry concentration. Still, it was hard to give any definitive interpretation of such regressions. Imagine that large firms have high fixed costs and low marginal costs, and low marginal costs are associated with higher markups (in part because the price needs to recover the high fixed costs). This can create a correlation between firm size or the Herfindahl–Hirschman index for an industry and markups.

One way of approaching the Demsetz (1973) empirical critique is that concentration is econometrically endogenous, suggesting a search for possible instruments. However, in many cases it is not at all clear what variables are excluded from the "concentration-markup" regression, which naturally depends on all elements of demand and marginal cost.

However, the critique runs deeper than concerns over endogeneity. Different changes in primitives, with very different positive and normative implications, can produce the same observed correlations between concentration and markups. Demsetz (1973) emphasized the path from improved marginal cost to the joint outcome of concentration and measured accounting markups. This path can exist even in a model of perfect competition with heterogeneous upward-sloping marginal cost curves. In contrast, the original structure-conduct-performance researchers emphasized the path from exogenous mergers to the joint outcome of high concentration, higher prices, and reduced consumer welfare, which offers an equally coherent story. One can also tell a story in a differentiated products context, in which a reduction in search or trade costs may shift market share toward firms with high-quality products, increasing both concentration and consumer welfare (as emphasized in Autor et al. 2017).

In short, there is no well-defined "causal effect of concentration on price," but rather a set of hypotheses that can explain observed correlations of the joint outcomes of price, measured markups, market share, and concentration.[1] As Bresnahan (1989) argued three decades ago, no clear interpretation of the impact of concentration is possible without a clear focus on equilibrium oligopoly demand and "supply," where supply includes the list of the marginal cost functions of the firms and the nature of oligopoly competition.

Some of the recent literature on concentration, profits, and markups has simply reasserted the relevance of the old-style structure-conduct-performance correlations. For economists trained in subfields outside industrial organization, such correlations can be attractive. Our own view, based on the well-established mainstream wisdom in the field of industrial organization for several decades, is that regressions of market outcomes on measures of industry structure like the Herfindahl–Hirschman index should be given little weight in policy debates. Such correlations will not produce information about the causal estimates that policy demands. It is these causal relationships that will help us understand what, if anything, may be causing markups to rise.

Detailed Industry Studies of Market Power

What kind of studies might provide better-grounded evidence on the underlying causes of shifts in concentration or markups?

As a starting point, we might seek to establish a descriptive baseline for analysis, without jumping to causal statements. Is concentration in general rising across many firms and industries or a relatively small number? Are accounting markups rising? Are prices rising? What are the descriptive correlations across these variables? The answers to these questions can often point to fruitful areas for detailed study as well as rule out concerns that are unsupported by the facts. We can then consider approaches to interpreting these fact patterns that may lead us to firmer policy conclusions.

[1] As a more specific example, in the Cournot model, the Lerner index of price-cost markups is equal to the Herfindahl–Hirschman index divided by the absolute value of the market demand elasticity (Cowling and Waterson 1976). If we could somehow empirically identify an industry-specific coefficient on the Herfindahl–Hirschman index in a regression of the correctly measured Lerner index on concentration, we would learn only one demand parameter, not nearly enough to know (for example) how a merger would affect industry markups. Even within the Cournot model, reductions in marginal cost will produce one kind of joint effect on the Herfindahl–Hirschman index and markups, whereas a merger will produce an altogether different set of joint effects (Farrell and Shapiro 1990). Most industries are, of course, not well approximated by the Cournot model, and extracting causal predictions from those industries is even harder.

As an example, Ganapati (2018a) builds on and extends recent work to address some of these correlational issues. In common with other authors, he finds a rising economy-wide trend toward increased concentration. Using industry-level price indices, in a difference-in-difference analysis he finds that "concentration increases are positively correlated to productivity and real output growth, uncorrelated with price changes and overall payroll, and negatively correlated with labor's revenue share." Autor et al. (2017) use firm-level panel data to document that the increase in concentration is largely due to reallocation of market share toward the preexisting set of large and productive firms. This change is associated with a decrease in the labor share. They provide a model that attributes these correlations to the rise of "superstar" productive firms. Although a number of authors report findings of increasing concentration across a wide range of industries, this finding is not universal. For example, Rossi-Hansberg, Sarte, and Trachter (2019) find falling concentration in local product markets, in part because entry of national firms will increase competition in local markets.

As an alternative, there has been a recent wave of "production function" approaches to measuring markups. These studies often use data from the financial accounts of firms to estimate firm-level production functions, which in turn serve as a basis to estimate the size of markups. One advantage of this approach is that it directly addresses the issue of markups in the economy as a whole. Another advantage is that these papers do not use measures of industry concentration, and thus they do not suffer from the fundamental methodological flaws of papers that use the structure-conduct-performance paradigm. However, a corresponding disadvantage of broad-based approaches to estimating markups by using financial accounting data or aggregate data is that modeling and estimation approaches that fail to model industry-specific characteristics restrict the range of answers that we can learn from data. We believe that this research provides persuasive evidence that markups have been rising, although open questions remain about the magnitude and causes of the effect. In this symposium, the articles by Susanto Basu and by Chad Syverson discuss this approach in detail.[2]

However, the main focus of this article is to discuss what we can conclude from industry-specific studies about the sizes and causes of markups and therefore what policy responses would be appropriate. In these industry-level studies, it may be plausible to identify markups from data on prices and output, together with data

[2] Prominent examples of this production function approach with US data include De Loecker and Eeckhout (2017, 2018a); Hall (2018); and Eggertsson, Robbins, and Getz Wold (2018). For example, De Loecker and Eeckhout (2017) in their primary analysis use firm-level financial statements from Compustat, including measures of sales, spending on inputs, capital stock, and industry classifications. Studies using this general approach on international data include De Loecker and Eeckhout (2018a) and Calligaris, Criscuolo, and Marcolin (2018). All of these papers find evidence of positive and rising markups. These studies show not just that markups are rising overall, but the fact that the rise in markups is due to a small number of firms. Again, for additional details, see the articles by Basu and by Syverson in this symposium. For other careful discussions, see also Yurukoglu (2018) and Raval (2019), as well as De Loecker and Eeckhout (2018b) for a response to criticisms.

on demand and cost shifters and some industry-appropriate assumptions about competitive behavior. Detailed industry studies can provide direct evidence on the causes and consequences of imperfect competition. The relatively narrow focus of industry-specific studies may frustrate economists who are accustomed to working with all firms in one model and dataset, as is often the case in macroeconomics and finance. But the nature of the demand, costs, and competitive setting that affect firm choices is inherently heterogeneous.

Here, we do not try to review the vast literature in this area, but instead focus on a few recent studies that illustrate some contexts in which this research is done and how the welfare implications of such research can be ambiguous, combining elements of lower cost, improved quality, and decreased competition.

As a first example, Ganapati (2018b) studies the large wholesaling sector of the economy. Ganapati notes that, in 2012, wholesalers accounted for 50 percent of sales to downstream buyers in the US manufactured goods market and that, contrary to prominent examples of large retailers disintermediating wholesalers, the wholesale sector overall was growing in size. As the wholesale sector has grown, it has become more concentrated, and accounting markups have increased. This has happened largely due to increases in the market shares of the largest wholesalers. This increase in concentration has been accompanied by increased spending on information technology, by the opening of warehouses closer to consumers, and by increased dual sourcing from domestic and foreign sources. Purely from the descriptive data, this story seems more complicated than either "perfect competition" or a classic Cournot-style oligopoly story of increased homogeneous goods concentration leading to higher prices and reduced output.

To interpret these trends, Ganapati (2018a) applies a series of standard empirical industrial organization models of demand, pricing, and entry. These models are fitted to detailed US Census data, with identification coming from "supply and demand"-style instrumental variable methods (Berry and Haile 2014). In particular, he uses data on the number of wholesalers by type and location, on market size, and on shifters of marginal cost. Ganapati concludes that the growth in the wholesale sector is driven by a combination of lower marginal costs and increased demand, which is in turn driven by an improved warehouse network as well as improved sourcing quality from both domestic and foreign locations.

The benefits of these improvements for downstream customers are constrained by lessened competition that yields an increase in markups over marginal cost. In Ganapati's (2018a) entry model, improved product quality and lower marginal costs are associated with higher fixed costs that are created by the firm's location, quality, and sourcing decisions (similar to the "endogenous fixed cost" models of Sutton 1991). However, Ganapati does not attempt to attribute these fixed costs to any specific source. They could be the information technology costs of improved logistics or the sunk costs of building out a warehouse structure. Alternatively, they could represent a rent due to oligopolistic behavior and (perhaps) first-mover advantages in establishing wholesale networks. The findings indicate that in this sector, while concentration and markups are rising, quality is rising and costs are falling, thus leading to a setting that is not easy to evaluate. Research on a number of other prominent industries finds patterns with similarly ambiguous welfare implications.

Note that, unlike in the structure-conduct-performance or the production function approaches mentioned above, Ganapati is able to make statements about demand, marginal costs, and fixed costs. While these statements depend on a significant number of maintained assumptions, they lead to a rich story about the underlying forces behind markup changes, and they lead to both positive and normative implications associated with those changes. Ganapati's work on wholesaling reveals an evolving industry with endogenous trade-offs in product quality, marginal costs, and fixed costs.

The airline industry provides another example in which increasing markups are associated with some degree of product improvement and marginal cost decline (Berry 1990), but it also illustrates that poorly policed mergers can increase prices. Debates over airline mergers often pivot on the negative effects of increased markups on some concentrated nonstop routes versus the potential for improved route structures leading to better choices and increased competition on other (often connecting) itineraries.[3] Borenstein (1990) notes the strong evidence that prices rose after at least two Reagan-era mergers of airlines with largely overlapping route networks. A more recent airline merger wave has consolidated the remaining legacy carriers into three large firms that face competition from Southwest Airlines and a group of new, low-cost carriers. We await a full academic evaluation of these mergers. The many years of near-zero-profit operations of major airlines (Borenstein 2011), lasting up until the demand boom and merger wave of recent years, suggest that for a long time, high markups over marginal cost in the industry were offset by the costs of running large hub-and-spoke networks. These networks create large benefits by providing low-priced and convenient connections through hubs to many destinations. But they also have allowed airlines to charge high markups on many direct flights out of hub airports (Berry, Carnall, and Spiller 2006).

Airlines, then, provide a rich but mixed example of the sources of markups. Running a hub-and-spoke network does involve endogenous fixed and sunk costs, but the possible effects of mergers on prices suggest a large role for antitrust policy in reducing harmful effects on consumers. The firms that provide local cable television and internet broadband may offer another example of monopoly rents (from deregulated physical connections at the household level) plus improved product quality (from new channels and increased speed), with markups protected in large part by the high fixed cost of adding new wired connections at the household level. It may well be that consumer surplus (and "output") is increasing in this industry, but not by as much as it would under alternative regulatory structures.

In other industry studies, higher concentration and markups do not seem to be accompanied by any improvement in quality. For example, many studies have shown that hospital consolidation between close competitors leads to substantial

[3] These debates follow the emphasis on improved airline product quality in Carlton, Landes, and Posner (1980) versus the emphasis on airline market power in Borenstein (1990).

increases in price and markups without improving quality (for example, Town and Vistnes 2001; Capps, Dranove, and Satterthwaite 2003; Gowrisankaran, Nevo, and Town 2015; Ho and Lee 2017) or leads to reductions in quality in price-regulated markets such as Medicare or the English National Health Service (Kessler and McClellan 2000; Cooper et al. 2011; Gaynor, Moreno-Serra, and Propper 2013; Gaynor, Propper, and Seiler 2016). For an overall review of this literature, see Gaynor, Ho, and Town (2015). With the exception of the associations identified by Cooper at al. (2019), research has not focused on identifying the major industry-wide factors driving higher hospital prices or markups. There has been little work examining entry or recovery of fixed costs (for an exception, see Abraham, Gaynor, and Vogt 2007) or whether fixed costs are rising. Moreover, it should be noted that separately identifying costs and rents is a challenge in the hospital industry. Many hospitals (particularly the largest) are not-for-profit; thus, rents tend to be spent and to appear as expenses (as is true for not-for-profit firms in general). Identifying and understanding the major factors driving increased hospital markups constitute a key next step in understanding this market.

A final issue is that when markups are measured as a ratio of prices to marginal costs, the rise in markups may be driven by very low marginal costs, as in a number of media and internet markets. For example, Waldfogel (2015) documents that in the recorded music industry, digitization lowered marginal distribution costs and the fixed costs of production, although “quality” is still produced via endogenous fixed costs. These lower costs led to an explosion of product variety. In such media and internet information markets, the “macro-production markup,” measured as the ratio of price to marginal cost, may go to near infinity as the marginal cost of the product declines to near zero, as long as the price remains clearly positive. Similarly, monopsony power can in principle also be a driver of increased markups via reduced marginal costs.

We have provided examples of three kinds of results from detailed industry studies. In some cases, such as wholesaling, investments may be generating product quality improvement together with a shift from marginal to fixed costs, yielding an improvement in consumer welfare. In other industries, such as airlines, markups may be associated with some quality improvement, but some mergers have also clearly resulted in price increases. In other markets, such as hospitals, there is no evidence that consolidation is resulting in systematic product quality improvements or clear cost reductions, but there is strong evidence of price increases (or quality reductions). The diversity of results across these industries is evidence of the value and richness that can be obtained from careful industry studies. It also serves as a caution of the difficulties of drawing useful inferences from aggregate studies across industries.

Industrial organization industry studies, taken as a whole, do provide evidence against some particularly simple or stylized models. These studies clearly reject models that would closely approximate perfect competition. Similarly, these studies emphasize important game-theoretic oligopoly features of markets, rejecting simple interpretations associated with the “Chicago School” of antitrust (for example, Bork 1978).

Instead, these industrial organization studies also suggest a nuanced reality in which large firms are in fact changing products and production methods, including the mix of marginal and fixed costs, over time. The industry studies seem to suggest that "fixed costs" are often actually sunk costs that are built up through time via investments in networks, product quality, geographic location, and so forth. An interesting question is how this possible reallocation from marginal to fixed costs affects labor demand. Another important question is whether the share of labor in variable costs is higher or lower than the share of labor in fixed costs.[4]

Of course, the discussion here covers just a small collection of industry studies. In our view, industry-level studies are required to understand the forces shaping markets in the modern economy and thereby to craft appropriate policies. These studies will have to take on broader segments of the overall economy if they are to fully respond to questions about aggregate markup trends. Also, while many existing industrial organization industry-level studies provide information on the level of markups, we would welcome a surge of industry-level research focused on trends in markups in order to discover where they are rising and why. By their nature, detailed industry studies will tend to produce estimates and explanations for markups that are more complex than those advanced in studies making use of broad-based financial accounting data or Census data aggregated across large numbers of firms in very different industries. Focusing at the industry level allows researchers to study the ways in which firms seek to create competitive advantages with a mixture of strategies, including investment in fixed capital, changes in product quality, geographic advantage, and consolidation by merger.

Factors Leading to Rising Markups

It seems plausible that some of the primitives of modern industrial organization—cost conditions, demand conditions, and pricing environment—have been changing over the past few decades. For example, the adoption of information technology is often a fixed cost involving hardware, such as servers, or software, such as enterprise resource planning software. Thus, firms and industries for which information technology has grown in importance have rising fixed costs, which leads to rising markups and can lead to markets dominated by one or a small number of large firms. On the demand side, the growing importance of network effects can lead to one or a small number of firms dominating a market and thus commanding

[4] As a contrast with this portrayal of evolving industries, a number of studies of markups are based on stronger assumptions. As one example, consider the (intentionally) highly stylized model of Autor et al. (2017). In that model, firms exogenously differ in their Hicks-neutral productivity shocks. There is a fixed labor requirement, common to all firms, which explains the negative correlation between firm size and the labor share. Motivated by the results from their firm-level production-side data, they then state that changes in industry average markups over time are explained by a reallocation of market share (as through lower trade or search costs). As more consumers purchase from the largest firms, the fixed labor requirement is spread over yet more units, raising markups still further.

higher markups. With regard to firm conduct, increased managerial exploitation of market power can lead to rising markups, as can the documented slow decline in US antitrust enforcement (for example, Baker 2019). In this section, we consider the available evidence on the factors that have been leading to rising markups.

Rising Fixed and Sunk Costs

We have already mentioned the models of Shaked and Sutton (1982) and Sutton (1991), where fixed (and often sunk) costs at the firm level partly reflect endogenous choices of product quality, production techniques, and marketing. Under the assumptions of these models, industries do not deconcentrate even as market size grows because there is always an incentive for some firm to become large, relative to the market, by making a sunk investment that drives up demand for its product.

Sutton (1991) gives examples where the better product does not involve much higher marginal cost (or can even involve reduced marginal cost), and therefore competition from lower-quality competitors does not compete away the markup of the firm producing the high-quality product. He argues that, during the period from the late nineteenth to the mid-twentieth century, decreasing transportation costs and national marketing strategies allowed many consumer goods products to trade higher fixed costs for national sales dominance. These firms maintained high markups and high national market shares in the absence of important scale economies of production. If Census data on production had existed during that period, they might have revealed a trend of increasing markups in consumer goods markets, with much of the markup attributable to a small number of "superstar" products.

What changes in the past few decades might allow firms to pursue a similar strategy of higher fixed costs and sustained market dominance? If a rise in the quality of services can be achieved with higher spending on information technology, and if a large component of information technology spending represents fixed costs, then the proportion of fixed to variable cost will be rising across the decades of increasing technological advancement. For example, Bessen (2017) provides evidence that customized software—used routinely by large corporations today—requires large up-front fixed sunk costs. Calligaris, Criscuolo, and Marcolin (2018) find higher markups in more digitally intensive industries and that differences in markups between digitally intensive and nonintensive industries have grown.

These patterns are consistent with the hypothesis that rising fixed sunk costs and lower marginal costs due to increases in information technology investments could be a significant driver of increasing markups. In studying this hypothesis, how can researchers measure fixed and sunk costs? As noted, industrial organization economists have often been suspicious of attempts to directly measure fixed costs from accounting or Census data because accounting rules do not follow economic principles for expensing, depreciation, rents on existing assets, and so forth.[5] Thus,

[5] This point is related to arguments in Fisher and McGowan (1983) and Schmalensee (1989) about general problems with depreciation, accounting data, and measured components of profit and cost.

industry-level studies typically estimate fixed (or sunk) costs as a kind of residual that explains the observed equilibrium market structure (or pattern of entry and exit; see Bresnahan and Reiss 1990; Berry 1992; Ciliberto and Tamer 2009; Berry, Eizenberg, and Waldfogel 2016). Fixed costs are bounded above by the level that would render existing firms unprofitable and below by the level that would induce incremental entry.

However, this approach treats fixed costs as exogenous. In some instances, a firm can choose its fixed costs, such as its level of advertising and promotion or of research and development. Treating fixed costs as endogenous is also consistent with evidence for the increased importance of intangible assets, which include management effectiveness, business processes, intellectual property, branding, and the effective use of information technology, as documented by Corrado, Hulten, and Sichel (2009), Haskel and Westlake (2017), and Bhandari and McGrattan (2018). Firms' market shares are positively correlated with their intangible assets, as Crouzet and Eberly (2018) demonstrate. Moreover, they show that in some sectors, such as consumer goods, higher intangible assets are positively correlated with higher productivity, while in other sectors, such as health care, intangible assets are correlated with higher measured markups. A rising role for intangible assets will further complicate the use of accounting data to discuss markups, since these assets may be treated in an inconsistent fashion in accounting data (Yurokoglu 2018).

The welfare consequences of increasing sunk and fixed costs in an industry are complex, are probably industry specific, and may vary across antitrust and regulatory regimes. On the consumer side, higher fixed costs may enable a rise in product quality, which is generally good. However, fixed costs may be duplicated by competitors, such that oligopoly generates excessive entry from the social welfare perspective (Mankiw and Whinston 1986; Berry and Waldfogel 1999). Moreover, better products may contribute to higher markups, especially if the high fixed (or sunk) costs limit the number of competing firms and drive up prices. Alternatively, higher markups can reflect falling marginal costs rather than higher prices.

On the firm side, fixed costs must be offset by positive markups in order for the firm to survive. Therefore, industries with high markups may or may not be profitable. Profits in excess of those necessary to cover current fixed costs might reflect a return on past investments; indeed, the expectation of a current stream of profits may have been necessary to bring forth a socially valuable innovation. In other cases, current profits may reflect a rent on past luck or may result from a past sunk investment that is preventing socially desirable entry (for the modern game theory of sunk costs and entry barriers, see Tirole 1988). It is difficult to see how cross-industry studies can capture the industry-level complexity that results from high fixed and sunk costs.

The distributional consequences of higher fixed costs, perhaps combined with lower marginal costs, can be equally complex. For example, it is easy to imagine cases where labor is particularly associated with variable product costs, while (for example) fixed costs are associated with the employment of software engineers and with returns to various forms of intellectual property. In some cases, imputed fixed costs may reflect rents that do not serve an efficiency-enhancing purpose. For example, one possible rent involves a return to a (possibly lucky) first-mover advantage in a network industry, as we discuss in the next subsection.

In our opinion, both industry studies and accounting data studies point to the broad category of endogenously increasing fixed and sunk costs as an important, perhaps the most important, source of the apparent pattern of rising global markups. In the next section, we focus on the specific case of network effects, which create particular complexities.

Network Effects

Network effects have become important in many sectors of the economy. In particular, they are often strongly present in digital platforms (US Bureau of Economic Analysis 2018), where many consumers rely on platforms with user-provided content regarding restaurants, hotels, traffic, and news. Network effects lead to markets dominated by one or a small number of firms, as in social media.

A rising importance of network effects can lead to weaker competition and thus higher markups in various ways. First, network effects tend to lead to consumer lock-in, enhancing firms' short-run market power while making new entry difficult. Second, network effects can make fixed costs more important, including expansions of information technology, distribution, delivery, and promotion in order to reach a larger number of customers. Third, the aggregation of eyeballs and consumer information by platforms may give an advantage to the dominant business in selling advertising and thus may perpetuate a concentrated market structure (Bergemann and Bonatti 2018). For these reasons, the locus of competition in network markets often turns out to be *for* the market, not *in* the market. Once a firm has come to dominate a network market, its market position is not easily eroded.

The lucky first mover in a market with network effects will benefit from these effects. Thus, markups in this instance include a rent on that luck, and there is no reason to believe that the (expected) market rent was required to generate the initial investment effort. Of course, the network can also create substantial consumer surplus. The policy question is whether some alternative antitrust or regulatory structure could improve the market outcome while retaining the consumer benefits.

Growing Monopsony Power

Claims have been made that the concentration of employers is growing in labor markets and that more concentrated employer markets are associated with lower wages (Azar, Marinescu, and Steinbaum 2017; Azar et al. 2018; Posner, Weyl, and Naidu 2018).[6] To the extent that these forces trended toward more monopsony power or more exercise of monopsony power over recent decades, the declining cost of labor, typically a variable cost, may have contributed to the trend in markups.

[6] The finding is not universal. Lipsius (2018) and Rinz (2018) find that employer concentration has fallen, implying that monopsony power has fallen, not risen.

There is long-standing evidence of monopsony power in some labor markets, notably the markets for nurses (Sullivan 1989; Currie, Farsi, and MacLeod 2005; Staiger, Spetz, and Phibbs 2010), teachers (Ransom and Sims 2010), and fast-food workers (Card and Krueger 1994). However, there is evidence that the extent of monopsony power in the labor market has grown over the years (Manning 2003). Some possible reasons include declines in union membership, in the powers available to unions, and in legal remedies available to individual workers—all of which have weakened worker bargaining power (Farber et al. 2018). There is also some evidence of the use of outsourcing by firms ("fissuring") to facilitate wage discrimination in a way that leads to lower average wages and higher markups (Weil 2011). There is speculation that the rise of the "gig" economy may be holding down worker wages as well (Dube and Kaplan 2010; Chen et al. 2017). Another feature of labor markets that likely grew over past decades but has been uncovered only recently is the use of noncompete clauses by employers in some industries (Starr, Prescott, and Bishara 2019), particularly for low-wage workers in fast-food and other franchises (Krueger and Ashenfelter 2018).[7]

A main difficulty in this area is that most of the existing studies of monopsony and wages follow the structure-conduct-performance paradigm; that is, they argue that greater concentration of employers can be applied to labor markets and then proceed to estimate regressions of wages on measures of concentration. For the same reasons we discussed above, studies like this may provide some interesting descriptions of concentration and wages but are not ultimately informative about whether monopsony power has grown and is depressing wages.

Recently, efforts have been made to take a sounder empirical approach. Card et al. (2018) review the evidence on labor markets and reconcile a variety of empirical results via a model of "differentiated jobs" that recalls industrial organization models of differentiated products. Azar, Berry, and Marinescu (2019) estimate an industrial organization–style model of differentiated job vacancy demand at the level of the job applicant applying for a specific job title within a commuting zone. They find moderately positive levels of firm market power even in labor markets that are not highly concentrated. However, this work estimates levels of labor market power, not trends over time.

Linkages can also arise between mergers and increased monopsony power. Prager and Schmitt (2019) examine the effect of mergers in the hospital industry and find evidence that mergers between nearby hospitals depress wage growth for workers with hospital-job-specific skills (but not for workers with general job market skills).

[7] The Washington State attorney general has challenged these noncompete agreements and by 2019 had achieved many dozens of settlements to not enforce and to remove the provisions. Also, the US Department of Justice has recently prosecuted multiple cases of firms explicitly agreeing not to hire away each other's workers (the "no poach" agreements), as well as naked collusion to fix wages that occurred over many years. One of the first of this recent group of cases involved many of the top employers among the Silicon Valley tech firms such as Apple, Google, Adobe, Intel, Intuit, and Pixar (*In re: High-Tech Employee Antitrust Litigation*, N.D. Cal. Case 11-CV-02509-LHK [2015]).

At present, the extent to which any decreased competition in the labor market is a major driver of increased markups is not clear, and research that sheds light on this question would be most welcome.

Increased Rent Seeking

Yet another potential explanation for higher markups is that managers are increasingly better trained (perhaps in economics or MBA programs) to find and exploit situations where their firms face inelastic demand. Firms in many industries, including airlines, entertainment, and retail, have improved over time in their ability to price discriminate, presumably raising some markups while lowering others, with an uncertain implication for the distribution of markups. Traditionally, the economics profession has treated these situations as arbitrage of informational rents that guide economic activity and lead to an increase in efficiency (an idea attributed to Friedrich von Hayek). But once exposed to public scrutiny, these instances are often portrayed and perceived as exploitation of consumers.

Some firms have gone beyond more aggressive price discrimination and have raised prices by engaging in holdup of a relationship-specific investment or by reneging on agreements that are not sufficiently protected by contract. In one example, pharmaceutical industry CEO Martin Shkreli sharply increased the price of a generic drug in a marketplace where it takes several years for a competitor to be approved by the Food and Drug Administration (Pollack 2015). In another example, holders of standard essential patents demanded high royalties from handset makers after networks implementing the standard were fully built out and could not be changed (Scott Morton and Shapiro 2016). In yet another example, hedge funds bought up the television stations that were needed to re-pack spectrum, so it could be used by wireless carriers, and strategically withheld those stations to raise the price of their assets (Doraszelski et al. 2017). And physicians who are out-of-network with a certain insurer charge patients in the in-network hospital where they work three times as much as in-network physicians would charge (for an example of out-of-network billing for emergency care, see Cooper, Scott Morton, and Shekita 2017). When one of the outsourcing companies that perfected this strategy was written up in the *New York Times* and the strategy became public (Creswell, Abelson, and Sanger-Katz 2017), insurers used the subsequent call for regulation to improve their bargaining positions in new contracts, and the outsourcing company's profits fell.

To the extent that firms and their managers are becoming more sophisticated in their pursuit of inelastic niches where they can create and exploit market power, the relevant markups will rise. Research that sheds some light on the extent of this phenomenon, whether it has grown, and whether and to what extent it has contributed to increased markups would be beneficial.

Globalization

Although globalization is not our focus here, it may also be part of the explanation of rising markups for the highest-markup firms. A market that contains some firms that globalize and others that do not could generate this pattern. Firms with a global supply chain will have access to lower-cost inputs and may then achieve economies of scale, leading to a higher markup. If such a globalized firm gains market share at the expense of domestic rivals, industry markups will rise. Thus, increased globalization may play a role in both increasing markups and the unequal distribution of the increase. Uncovering what effects globalization may have had on markets and markups seems a potentially fruitful area for future research.

Antitrust Enforcement

There were undoubtedly some cases of overly aggressive enforcement of antitrust laws in the 1960s and 1970s; in one much-discussed case, courts upheld blocking a merger that would have resulted in a combined market share of 7.5 percent (*United States v. Von's Grocery Company*, 384 US 270 [1966]). However, courts in recent decades have been steadily dialing back antitrust enforcement, both through economic assumptions built in to jurisprudence and through practical changes such as raising the pleading standards for plaintiffs (Baker 2019; Gavil 2019). Mergers in markets with more than two firms are much less likely to be challenged now than in past decades (Kwoka 2016). The recent *Ohio v. American Express Company* (138 S. Ct. 2274 [2018]) Supreme Court ruling has been interpreted by some as possibly ending the government's ability to bring an antitrust case against a platform that operates in a two-sided market (Open Markets 2018).

The decline of antitrust enforcement in recent decades may be a contributor to rising markups, although more research is needed to substantiate this conclusion firmly (Kulick 2017; Baker 2019; Wollmann 2019). However, antitrust enforcement and competition policy is important in this context because, unlike shifts in fixed costs and technology, it can be directly addressed with legislation. Moreover, regardless of the role of changing antitrust enforcement in explaining a rise in markups, higher markups imply a world that may require increased antitrust vigilance.

Here, we provide an overview of some commonly mentioned concerns about underenforcement of antitrust laws that are especially applicable to the large, high-markup firms most at issue: vertical restraints, coordinated effects, digital platforms, exploitation of intellectual property, acquisition of potential competitors, and exclusionary conduct. These issues have been discussed in more detail in a number of policy venues (Baker 2019; Scott Morton et al. 2019; Federico, Scott Morton, and Shapiro forthcoming; Shapiro in this issue). We then offer some concluding thoughts on the appropriate perspective of antitrust enforcement given the current state of knowledge in these areas.

Some Specific Concerns about Underenforcement of Antitrust Laws

The term *vertical restraints* describes contracts between firms with a vertical relationship that may have anticompetitive effects depending on the type of restraint, the party using it, market structure, and so forth (Segal and Whinston 2000; Conlon and Mortimer 2013; Asker 2016; Crawford et al. 2018). These issues seem potentially important in the current situation where certain markets have come to be dominated by one or a small number of large firms. A common situation is that high-markup platform firms succeed by offering valuable (often digital) goods and services to consumers, but then competition issues arise when the platform either begins to supply the complementary products itself or contracts over price, quality, or technology in a way that limits the independent complements on the platform. Raising rivals' costs, foreclosure, and exclusion are among the possible theories of harm that can be raised in this setting. The Vertical Merger Guidelines of the US Department of Justice were last updated in 1984, and the federal agencies rarely bring such cases. The government litigated its first vertical merger case in 40 years in 2018, arguing that the proposed vertical merger between AT&T and Time Warner was anticompetitive, but lost convincingly at the federal appeals court level (*United States v. AT&T Inc., DirecTV Group Holdings, LLC, and Time Warner Inc.*, 310 F. Supp. 3d 161 [2018]).

The term *coordinated effects* refers to a situation in which concentrated industries or sectors may be more susceptible to tacit collusion (Tirole 1988). Recent empirical work has found tacit collusion to be unexpectedly prevalent (Ciliberto and Williams 2014; Miller and Weinberg 2017; Schmitt 2018), but in general, the economics profession has contributed little to this policy area. In a world with trends toward concentration, more understanding and measurement of tacit collusion would be valuable.

The rise of *digital platforms* has been an important change in the economy, sparking rising calls from some quarters for antitrust action against firms such as Amazon, Facebook, and Google (Khan 2017; Wu 2018; Hughes 2019). The European Commission has been active in this area, raising issues that include allegations of exclusionary bundling, anticompetitive exclusive contracts, vertical foreclosure, and anticompetitive mergers. In our view, establishing robust theories of harm and tools to evaluate the evidence for or against digital platforms is a valuable activity for the antitrust agencies as well as academic economists. However, US antitrust agencies have not been active in this area, with the exception of the investigation by the Federal Trade Commission that led to a settlement but no case (US Federal Trade Commission 2013).

Firms may *exploit intellectual property* by using patents or other intellectual property to engage in exclusionary conduct in related markets. For example, branded drugs have long used patent litigation settlements as a way to pay generic rivals to stay out of the market (called "reverse payments" or "pay for delay"). It took 18 years from the time the Federal Trade Commission first identified this strategy to the time when the US Supreme Court ruled that it can, under certain conditions, be illegal (*FTC v. Actavis, Inc.*, 570 US 136 [2013]). Pharmaceutical firms have also used "patent thickets" and "product hopping" (for example, changing dosages or packaging) to prevent competitive entry or substitution. Patent litigation can be used as a strategy by firms with large portfolios to discourage investment and innovation or to partner with an incumbent firm to disadvantage rivals: as one example, the Federal Trade Commission successfully sued Qualcomm for such tactics involving a key semiconductor device used in smartphones (for background, see the case summary and links on the Federal Trade Commission's website at https://www.ftc.gov/enforcement/cases-proceedings/141-0199/qualcomm-inc). A similar result occurs when a standard-setting organization for an industry sets a standard that requires the use of an essential patent—and then the firm holding that patent denies rivals access to the patent on fair, reasonable, and nondiscriminatory terms. In work on causes behind a rise in dominant firms and a fall in US business dynamism, Akcigit and Ates (2019) suggest that one cause is "a heavy use of intellectual property protection by market leaders to limit the dissemination of knowledge."

Acquisition of potential competitors when they are still small can be a way for a dominant firm to improve quality or to fold a complement into its core product—or just to block a future potential entrant. Traditional antitrust enforcement has often focused on whether a merger led to an immediate significant increase in market share, not on how it affected potential or nascent competition. But when a market is subject to strong network effects, competition is *for* the market, and the possibility that the nascent entrant could contest the incumbent is an important source of competition. Frequently mentioned anecdotes include big tech companies' acquisitions of small firms in adjacent product markets, such as Facebook's acquisitions of Instagram and WhatsApp. In a study of the pharmaceutical industry, Cunningham, Ederer, and Ma (2018) conclude that about 6.4 percent of pharma acquisitions are "killer acquisitions," where the acquisition eliminates entry by a potential competitor. However, both the probability and the value of potential entry are uncertain, and research on identifying or measuring these effects in different settings would be extremely useful.

Exclusionary conduct arises when large incumbent firms with low marginal costs undertake activities that deter entry or disadvantage existing rivals. Two of the many possible examples of exclusionary conduct especially relevant in the current context include most-favored-nation contracts and refusals to deal.

Most-favored-nation (MFN) contracts (a term lifted from international trade treaties) specify that a seller must give the buyer who has such a contract as good a price as that seller gives to any other buyer. This may appear procompetitive. But notice that MFN contracts make price discounts more costly for the seller—any discount to any other buyer must also be provided to the buyer with the MFN contract. For example, imagine the firms interacting on a large digital platform, like hotels, agree to sign an MFN contract with the platform (Boik and Corts 2016; Baker and Scott Morton 2018). If a rival digital platform with a lower commission (say, 10 percent instead of 25 percent) enters and contracts with the same hotels, the hotel room must be priced as high on the low-margin platform as it is on the high-margin platform, and the lower-cost distribution channel may fail to gain traction. These practices have been challenged in Europe, but not in the United States (Mantovani, Piga, and Reggiani 2017).

Refusals to deal and foreclosure can be attempts to weaken competition. The European Commission's case against Google's search engine illustrates this issue (European Commission 2017). Suppose a provider of local service listings is a complement to general search; namely, a consumer can search on Google and find a Yelp page that holds the desired information. Displaying the Yelp page and letting consumers learn about it may allow Yelp to establish an independent relationship with consumers. The platform can use its rules to determine the display of organic results and the selection of ads shown, and in this way, it may be able to steer consumers away from such a complement. The platform could have a financial interest in doing so because of the risk that consumers learn to go straight to Yelp, reducing single-homing and the market power of the platform. This strategy might be even more attractive if the platform sells its own (vertically integrated) similar local search product and can divert revenues from local search advertising to itself by steering customers to its own product. (Or perhaps it could raise its rival's costs by requiring it to purchase an ad in order to obtain consumers.) Foreclosure strategies of this type can reduce competition in either the underlying platform market or, possibly, in competition among services provided on the platform.

As the economy becomes increasingly digital, possessing data can be another way to limit competition. For example, health-care systems often refuse or make it difficult to transmit patients' data to alternative health-care providers, with the explicit goal of retaining patients (Savage, Gaynor, and Adler-Milstein 2019). Anticompetitive use of data is another method of exclusion. The US Department of Justice recently settled a case against a large hospital system for employing clauses in its contracts with insurers that prevented insurers from providing patients information or incentives that would direct them to lower-cost or higher-value hospitals (*United States and the State of North Carolina v. Carolinas Healthcare System*; see US Department of Justice 2016). Another such case is being pursued by the attorney general in California (*People of the State of California Ex Rel. Xavier Becerra v. Sutter Health*; see California Department of Justice 2018).

Moving Forward with Antitrust Enforcement in a Situation of Uncertainty

Much of the evidence regarding rising markups seems to us plausible and worthy of further investigation, although uncertainty remains as to the most important causes. But this uncertainty should not imply inaction in antitrust policy (for a decision-theoretic approach to antitrust enforcement, see Baker 2015). We do know that competitive markets are generally beneficial for consumers. We also know that market power, once acquired, can be durable due to many of the economic and strategic issues discussed above. In particular, a substantial game-theoretic literature emphasizes the role of sunk costs in maintaining high markups (Tirole 1988). There are many examples in US economic history, including IBM and Microsoft, in which substantial market power persisted over decades.

Our view is that the policy focus should be on forms of antitrust enforcement that are robust to the magnitudes that future research on these issues may uncover. We believe that the most useful focus for antitrust enforcers around the globe should be on conditions of entry, including acquisitions by existing firms of recent or potential entrants, along with exclusionary conduct. Without rules to ensure there is competition on the merits, existing market power can be leveraged to create future market power and generate the durability that appears in the data. Consistent, vigorous antitrust enforcement is needed to ensure that concentration does not perpetuate itself because entry is not protected.

It's worth remembering that government agencies besides the antitrust authorities at the Federal Trade Commission and the US Department of Justice can have significant impacts on entry, market structure, and competition. For example, rules from the Food and Drug Administration hinder entry of biosimilar drugs. The Department of Health and Human Services permits higher fees to be charged for the same physician service if the service is provided in a doctor's office owned by a hospital and permits hospitals (but not doctors) to obtain substantial discounts on expensive drugs (like those for treating cancer) that are administered by physicians (the Section 340B program). These policies unintentionally encourage consolidation, since hospitals and physician practices can share the rents from these regulatory loopholes if the practices are owned by hospitals. Rules from the US Department of Transportation (2017) affect the transparency of airline fees. The US Patent and Trademark Office's decision to issue low-quality patents enables the activities of patent trolls. The Federal Communications Commission sets rules that give multichannel video programming distributors greater or lesser power to limit content provision by online video providers. At the state level, legislatures respond to the desires of incumbent car dealers by passing laws preventing the entry of new car brands into the state (*Tesla Motors, Inc. v. Johnson et al.*, W.D. Mich. Civil Action 16-cv-1158 [2017]; Gavil, Feinstein, and Gaynor 2014).

In summary, a wave of industry-level econometric studies will be needed to help us understand shifts in markups, the underlying causes, and more broadly how markets in our modern economy are functioning and evolving. Many of the likely causes of rising markups in this article involve economic shifts that do not have any direct policy response. But whatever the underlying cause and size of rising markups, promoting competition along the lines mentioned here seems to us to be, at present, the most appropriate policy response.

■ *We thank Al Klevorick, Carl Shapiro, and Chad Syverson, as well as Editor Enrico Moretti, Coeditor Gordon Hanson, and Managing Editor Timothy Taylor for helpful comments and suggestions that substantially improved the paper. The usual caveat applies.*

References

Abraham, Jean Marie, Martin Gaynor, and William B. Vogt. 2007. "Entry and Competition in Local Hospital Markets." *Journal of Industrial Economics* 55(2): 265–88.

Akcigit, Ufuk, and Sina T. Ates. 2019. "What Happened to U.S. Business Dynamism?" Becker Friedman Institute Working Paper. https://bfi.uchicago.edu/working-paper/what-happened-to-u-s-business-dynamism/.

Asker, John. 2016. "Diagnosing Foreclosure Due to Exclusive Dealing." *Journal of Industrial Economics* 64(3): 375–410.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2017. "The Fall of the Labor Share and the Rise of Superstar Firms." NBER Working Paper 23396.

Azar, José, Steven Berry, and Ioana Marinescu. 2019. "Estimating Labor Market Power." Unpublished.

Azar, José, Ioana Marinescu, and Marshall I. Steinbaum. 2017. "Labor Market Concentration." NBER Working Paper 24147.

Azar, José, Ioana Marinescu, Marshall I. Steinbaum, and Bledi Taska. 2018. "Concentration in US Labor Markets: Evidence from Online Vacancy Data." NBER Working Paper 24395.

Baker, Jonathan B. 2015. "Taking the Error Out of 'Error Cost' Analysis: What's Wrong with Antitrust's Right." *Antitrust Law Journal* 80(1): 1–38.

Baker, Jonathan B. 2019. *The Antitrust Paradigm: Restoring a Competitive Economy*. Cambridge, MA: Harvard University Press.

Baker, Jonathan B., and Fiona Scott Morton. 2018. "Antitrust Enforcement against Platform MFNs." *Yale Law Journal* 127(7): 1742–2203.

Barkai, Simcha. 2017. "Declining Labor and Capital Shares." https://www.london.edu/faculty-and-research/academic-research/d/declining-labor-and-capital-shares.

Benmelech, Efraim, Nittai Bergman, and Hyunseob Kim. 2018. "Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages?" NBER Working Paper 24307.

Bergemann, Dirk, and Alessandro Bonatti. 2018. "Markets for Information: An Introduction." Cowles Foundation Discussion Paper 2142. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3240310.

Berry, Steven. 1990. "Airport Presence as Product Differentiation." *American Economic Review* 80(2): 394–99.

Berry, Steven. 1992. "Estimation of a Model of Entry in the Airline Industry." *Econometrica* 60(4): 889–917.

Berry, Steven, Michael Carnall, and Pablo T. Spiller. 2006. "Airline Hubs: Costs, Markups and the Implications of Customer Heterogeneity," in *Competition Policy and Antitrust*, edited by Darin Lee, 183–214. Advances in Airline Economics 1. Bingley, UK: Emerald Group.

Berry, Steven, Alon Eizenberg, and Joel Waldfogel. 2016. "Optimal Product Variety in Radio Markets." *RAND Journal of Economics* 47(3): 463–97.

Berry, Steven, and Philip A. Haile. 2014. "Identification in Differentiated Products Markets Using Market Level Data." *Econometrica* 82(5): 1749–97.

Berry, Steven, and Joel Waldfogel. 1999. "Free Entry and Social Inefficiency in Radio Broadcasting." *RAND Journal of Economics* 30(3): 397–420.

Bessen, James E. 2017. "Information Technology and Industry Concentration." Boston University School of Law, Law and Economics Research Paper 17-41. https://ssrn.com/abstract=3044730.

Bhandari, Anmol, and Ellen R. McGrattan. 2018. "Sweat Equity in U.S. Private Business." NBER Working Paper 24520, Federal Reserve Bank of Minneapolis Staff Report 560.

Boik, Andre, and Kenneth S. Corts. 2016. "The Effects of Platform Most-Favored-Nation Clauses on Competition and Entry." *Journal of Law and Economics* 59(1): 105–34.

Borenstein, Severin. 1990. "Airline Mergers, Airport Dominance, and Market Power." *American Economic Review* 80(2): 400–404.

Borenstein, Severin. 2011. "Why Can't US Airlines Make Money?" *American Economic Review* 101(3): 233–37.

Bork, Robert H. 1978. *The Antitrust Paradox*. New York: Free Press.

Bresnahan, Timothy F. 1989. "Empirical Studies of Industries with Market Power." Chap. 17 in *Handbook of Industrial Organization*, vol. 2, edited by Richard Schmalensee and Robert Willig, 1011–57. Amsterdam: Elsevier.

Bresnahan, Timothy F., and Peter C. Reiss. 1990. "Entry in Monopoly Markets." *Review of Economic Studies* 57(4): 531–53.

California Department of Justice. 2018. "Attorney General Becerra Sues Sutter Health for Anti-competitive Practices That Increase Prices for California Families." Press Release, Office of the Attorney General, California Department of Justice, March 30, 2018. https://oag.ca.gov/news/press-releases/attorney-general-becerra-sues-sutter-health-anti-competitive-practices-increase.

Calligaris, Sara, Chiara Criscuolo, and Luca Marcolin. 2018. "Mark-Ups in the Digital Era." Organisation for Economic Co-operation and Development (OECD) Science, Technology and Industry Working Paper 2018/10.

Capps, Cory, David Dranove, and Mark Satterthwaite. 2003. "Competitition and Market Power in Option Demand Markets." *RAND Journal of Economics* 34(4): 737–63.

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. 2018. "Firms and Labor Market Inequality: Evidence and Some Theory." *Journal of Labor Economics* 36(S1): S13–70.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84(4): 772–93.

Carlton, Dennis W., William M. Landes, and Richard A. Posner. 1980. "Benefits and Costs of Airline Mergers: A Case Study." *Bell Journal of Economics* 11(1): 65–83.

Chen, M. Keith, Judith A. Chevalier, Peter E. Rossi, and Emily Oehlsen. 2017. "The Value of Flexible Work: Evidence from Uber Drivers." NBER Working Paper 23296.

Ciliberto, Federico, and Elie Tamer. 2009. "Market Structure and Multiple Equilibria in Airline Markets." *Econometrica* 77(6): 1791–828.

Ciliberto, Federico, and Jonathan W. Williams. 2014. "Does Multimarket Contact Facilitate Tacit Collusion? Inference on Conduct Parameters in the Airline Industry." *RAND Journal of Economics* 45(4): 764–91.

Conlon, Christopher T., and Julie Holland Mortimer. 2013. "Efficiency and Foreclosure Effects of Vertical Rebates: Empirical Evidence." NBER Working Paper 19709.

Cooper, Zack, Stuart V. Craig, Martin Gaynor, and John Van Reenen. 2019. "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured." *Quarterly Journal of Economics* 134(1): 51–107.

Cooper, Zack, Stephen Gibbons, Simon Jones, and Alistair McGuire. 2011. "Does Hospital Competition Save Lives? Evidence from the English NHS Patient Choice Reforms." *Economic Journal* 121(554): F228–60.

Cooper, Zack, Fiona Scott Morton, and Nathan Shekita. 2017. "Surprise! Out-of-Network Billing for Emergency Care in the United States." NBER Working Paper 23623.

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2009. "Intangible Capital and U.S. Economic Growth." *Review of Income and Wealth* 55(3): 661–85.

Cowling, Keith, and Michael Waterson. 1976. "Price–Cost Margins and Market Structure." *Economica* 43(171): 267–74.

Crawford, Gregory S., Robin S. Lee, Michael D. Whinston, and Ali Yurukoglu. 2018. "The Welfare Effects of Vertical Integration in Multichannel Television Markets." *Econometrica* 86(3): 891–954.

Creswell, Julie, Reed Abelson, and Margot Sanger-Katz. 2017. "The Company behind Many Surprise Emergency Room Bills." *New York Times*, July 24, 2017. https://www.nytimes.com/2017/07/24/upshot/the-company-behind-many-surprise-emergency-room-bills.html.

Crouzet, Nicolas, and Janice Eberly. 2018. "Intangibles, Investment, and Efficiency." *AEA Papers and Proceedings* 108(1): 426–31.

Cunningham, Colleen, Florian Ederer, and Song Ma. 2018. "Killer Acquisitions." https://ssrn.com/abstract=3241707.

Currie, Janet, Mehdi Farsi, and W. Bentley MacLeod. 2005. "Cut to the Bone? Hospital Takeovers and Nurse Employment Contracts." *Industrial and Labor Relations Review* 58(3): 471–93.

De Loecker, Jan, and Jan Eeckhout. 2017. "The Rise of Market Power and the Macroeconomic Implications." NBER Working Paper 23687.

De Loecker, Jan, and Jan Eeckhout. 2018a. "Global Market Power." NBER Working Paper 24768.

De Loecker, Jan, and Jan Eeckhout. 2018b. "Some Thoughts on the Debate about (Aggregate) Markup Measurement." http://www.janeeckhout.com/wp-content/uploads/Thoughts.pdf.

Demsetz, Harold. 1973. "Industry Structure, Market Rivalry, and Public Policy." *Journal of Law and Economics* 16(1): 1–9.

Doraszelski, Ulrich, Katja Seim, Michael Sinkinson, and Peichun Wang. 2017. "Ownership Concentration and Strategic Supply Reduction." NBER Working Paper 23034.

Dube, Arindrajit, and Ethan Kaplan. 2010. "Does Outsourcing Reduce Wages in the Low-Wage Service Occupations? Evidence from Janitors and Guards." *Industrial and Labor Relations Review* 63(2): 287–306.

Eggertsson, Gauti B., Jacob A. Robbins, and Ella Getz Wold. 2018. "Kaldor and Piketty's Facts: The Rise of Monopoly Power in the United States." NBER Working Paper 24287.

European Commission. 2017. "Google Search (Shopping)." Case AT.39740. http://ec.europa.eu/competition/elojade/isef/case_details.cfm?proc_code=1_39740.

Farber, Henry S., Daniel Herbst, Ilyana Kuziemko, and Suresh Naidu. 2018. "Unions and Inequality over the Twentieth Century: New Evidence from Survey Data." NBER Working Paper 24587.

Farrell, Joseph, and Carl Shapiro. 1990.

"Horizontal Mergers: An Equilibrium Analysis." *American Economic Review* 80(1): 107–26.

Federico, Giulio, Fiona Scott Morton, and Carl Shapiro. Forthcoming. "Antitrust and Innovation: Welcoming and Protecting Disruption." Chap. 4 in *Innovation Policy and the Economy*, vol. 20, edited by Josh Lerner and Scott Stern. Chicago: University of Chicago Press.

Fisher, Franklin M., and John J. McGowan. 1983. "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits." *American Economic Review* 73(1): 82–97.

Furman, Jason. 2015. "Business Investment in the United States: Facts, Explanations, Puzzles, and Policies." Technical Report, Council of Economic Advisers, Executive Office of the President of the United States.

Furman, Jason, and Peter Orszag. 2018. "A Firm-Level Perspective on the Role of Rents in the Rise in Inequality." Chap. 1 in *Toward a Just Society: Joseph Stiglitz and Twenty-First Century Economics*, edited by Martin Guzman, 19–47. New York: Columbia University Press.

Ganapati, Sharat. 2018a. "Oligopolies, Prices, Output, and Productivity." https://ssrn.com/abstract=3030966.

Ganapati, Sharat. 2018b. "The Modern Wholesaler: Global Sourcing, Domestic Distribution, and Scale Economies." https://www.tuck.dartmouth.edu/uploads/content/Ganapati_Wholesalers_2016_copy.pdf.

Gavil, Andrew I. 2019. "Crafting a Monopolization Law for Our Time." *Competitive Edge* (blog), Washington Center for Equitable Growth, March 27, 2019. https://equitablegrowth.org/competitive-edge-crafting-a-monopolization-law-for-our-time/.

Gavil, Andy, Debbie Feinstein, and Marty Gaynor. 2014. "Who Decides How Consumers Should Shop?" *Competition Matters* (blog), US Federal Trade Commission, April 24, 2014. https://www.ftc.gov/news-events/blogs/competition-matters/2014/04/who-decides-how-consumers-should-shop.

Gaynor, Martin, Kate Ho, and Robert J. Town. 2015. "The Industrial Organization of Health-Care Markets." *Journal of Economic Literature* 53(2): 235–84.

Gaynor, Martin, Rodrigo Moreno-Serra, and Carol Propper. 2013. "Death By Market Power: Reform, Competition, and Patient Outcomes in the National Health Service." *American Economic Journal: Economic Policy* 5(4): 134–66.

Gaynor, Martin, Carol Propper, and Stephan Seiler. 2016. "Free to Choose? Reform, Choice, and Consideration Sets in the English National Health Service." *American Economic Review* 106(11): 3521–57.

Gowrisankaran, Gautam, Aviv Nevo, and Robert Town. 2015. "Mergers When Prices Are Negotiated: Evidence from the Hospital Industry." *American Economic Review* 105(1): 172–203.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely. Forthcoming. "Are US Industries Becoming More Concentrated?" *Review of Finance*.

Gutiérrez, Germán, and Thomas Philippon. 2017a. "Declining Competition and Investment in the U.S." NBER Working Paper 23583.

Gutiérrez, Germán, and Thomas Philippon. 2017b. "Investment-less Growth: An Empirical Investigation." *Brookings Papers on Economic Activity* 2017(Fall): 89–169.

Hall, Robert E. 2018. "New Evidence on the Markup of Prices over Marginal Costs and the Role of Mega-firms in the US Economy." NBER Working Paper 24574.

Haskel, Jonathan, and Stian Westlake. 2017. *Capitalism without Capital: The Rise of the Intangible Economy*. Princeton, NJ: Princeton University Press.

Ho, Kate, and Robin S. Lee. 2017. "Insurer Competition in Health Care Markets." *Econometrica* 85(2): 379–417.

Hughes, Chris. 2019. "It's Time to Break Up Facebook." *New York Times*, May 9, 2019. https://www.nytimes.com/2019/05/09/opinion/sunday/chris-hughes-facebook-zuckerberg.html.

Kessler, Daniel P., and Mark B. McClellan. 2000. "Is Hospital Competition Socially Wasteful?" *Quarterly Journal of Economics* 115(2): 577–615.

Khan, Lina. 2017. "Amazon's Antitrust Paradox." *Yale Law Journal* 126(3): 710–805.

Krueger, Alan B., and Orley Ashenfelter. 2018. "Theory and Evidence on Employer Collusion in the Franchise Sector." NBER Working Paper 24831.

Kulick, Robert B. 2017. "Ready-to-Mix: Horizontal Mergers, Prices, and Productivity." US Census Bureau Center for Economic Studies Working Paper CES-WP-17-38. https://ssrn.com/abstract=2637961.

Kwoka, John E. Jr. 2016. "The Structural Presumption and the Safe Harbor in Merger Review: False Positives, or Unwarranted Concerns?" https://ssrn.com/abstract=2782152.

Lipsius, Ben. 2018. "Labor Market Concentration Does Not Explain the Falling Labor Share." https://ssrn.com/abstract=3279007.

Mankiw, N. Gregory, and Michael D. Whinston. 1986. "Free Entry and Social Inefficiency." *RAND Journal of Economics* 17(1): 48–58.

Manning, Alan. 2003. *Monopsony in Motion: Imperfect Competition in Labor Markets*. Princeton, NJ: Princeton University Press.

Mantovani, Andrea, Claudio A. Piga, and Carlo

Reggiani. 2017. "The Dynamics of Online Hotel Prices and the EU Booking.com Case." Networks, Electronic Commerce and Telecommunications (NET) Institute Working Paper 17-04. https://ssrn.com/abstract=3049339.

Miller, Nathan H., and Matthew C. Weinberg. 2017. "Understanding the Price Effects of the MillerCoors Joint Venture." *Econometrica* 85(6): 1763–91.

Open Markets. 2018. "Open Markets Statement on Ohio v. American Express Decision." Press Release, June 25, 2018. https://openmarketsinstitute.org/releases/open-markets-statement-ohio-v-american-express/.

Pollack, Andrew. 2015. "Drug Goes From $13.50 a Tablet to $750, Overnight." *New York Times*, September 20, 2015. https://www.nytimes.com/2015/09/21/business/a-huge-overnight-increase-in-a-drugs-price-raises-protests.html.

Posner, Eric A., Glen Weyl, and Suresh Naidu. 2018. "Antitrust Remedies for Labor Market Power." *Harvard Law Review* 132(2): 536–601.

Prager, Elena, and Matt Schmitt. 2019. "Employer Consolidation and Wages: Evidence from Hospitals." Washington Center for Equitable Growth Working Paper Series. https://equitablegrowth.org/working-papers/employer-consolidation-and-wages-evidence-from-hospitals/.

Ransom, Michael R., and David P. Sims. 2010. "Estimating the Firm's Labor Supply Curve in a 'New Monopsony' Framework: School Teachers in Missouri." *Journal of Labor Economics* 28(2): 331–55.

Raval, Devesh. 2019. "Testing the Production Approach to Markup Estimation." https://ssrn.com/abstract=3324849.

Ravenscraft, David J. 1983. "Structure–Profit Relationships at the Line of Business and Industry Level." *Review of Economics and Statistics* 65(1): 22–31.

Rinz, Kevin. 2018. "Labor Market Concentration, Earnings Inequality, and Earnings Mobility." US Census Bureau Center for Administrative Records Research and Applications Working Paper 2018-10. https://www.census.gov/library/working-papers/2018/adrm/carra-wp-2018-10.html.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter. 2019. "Diverging Trends in National and Local Concentration." https://www.princeton.edu/~erossi/DTNLC.pdf.

Savage, Lucia, Martin Gaynor, and Julia Adler-Milstein. 2019. "Digital Health Data and Information Sharing: A New Frontier for Health Care Competition?" *Antitrust Law Journal* 82(2): 592–621.

Schmalensee, Richard. 1989. "Inter-industry Studies of Structure and Performance." Chap. 16 in *Handbook of Industrial Organization*, vol. 2, edited by Richard Schmalensee and Robert Willig, 951–1009. Amsterdam: Elsevier.

Schmitt, Matt. 2018. "Multimarket Contact in the Hospital Industry." *American Economic Journal: Economic Policy* 10(3): 361–87.

Scott Morton, Fiona, Pascal Bouvier, Ariel Ezrachi, Bruno Jullien, Roberta Katz, Gene Kimmelman, A. Douglas Melamed, and Jamie Morgenstern. 2019. *Committee for the Study of Digital Platforms: Market Structure and Antitrust Subcommittee Report. Draft.* Chicago: Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business. https://research.chicagobooth.edu/-/media/research/stigler/pdfs/market-structure---report-as-of-15-may-2019.pdf.

Scott Morton, Fiona, and Carl Shapiro. 2016. "Patent Assertions: Are We Any Closer to Aligning Reward to Contribution?" Chap. 4 in *Innovation Policy and the Economy*, vol. 16, edited by Josh Lerner and Scott Stern, 89–133. Chicago: University of Chicago Press.

Segal, Ilya R., and Michael D. Whinston. 2000. "Naked Exclusion: Comment." *American Economic Review* 90(1): 296–309.

Shaked, Avner, and John Sutton. 1982. "Relaxing Price Competition through Product Differentiation." *Review of Economic Studies* 49(1): 3–13.

Smith, Noah. 2017. "America's Superstar Companies Are a Drag on Growth." *Bloomberg*, September 1, 2017. https://www.bloomberg.com/opinion/articles/2017-09-01/america-s-superstar-companies-are-a-drag-on-growth.

Staiger, Douglas O., Joanne Spetz, and Ciaran S. Phibbs. 2010. "Is There Monopsony in the Labor Market? Evidence from a Natural Experiment." *Journal of Labor Economics* 28(2): 211–36.

Starr, Evan, J. J. Prescott, and Norman Bishara. 2019. "Noncompetes in the U.S. Labor Force." University of Michigan Law and Economics Research Paper 18-013. https://ssrn.com/abstract=2625714.

Sullivan, Daniel. 1989. "Monopsony Power in the Market for Nurses." *Journal of Law and Economics* 32(2): S135–78.

Sutton, John. 1991. *Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration.* Cambridge, MA: MIT Press.

Tirole, Jean. 1988. *The Theory of Industrial Organization.* Cambridge, MA: MIT Press.

Town, Robert, and Gregory Vistnes. 2001. "Hospital Competition in HMO Networks." *Journal of Health Economics* 20(5): 733–53.

US Bureau of Economic Analysis. 2018. "Initial Estimates Show Digital Economy Accounted for 6.5 Percent of GDP in 2016." *BEA Wire*, March 15, 2018.

https://www.bea.gov/news/blog/2018-03-15/initial-estimates-show-digital-economy-accounted-65-percent-gdp-2016.

US Department of Justice. 2016. "Justice Department and North Carolina Sue Carolinas Healthcare System to Eliminate Unlawful Steering Restrictions." Press Release 16-665, Office of Public Affairs, US Department of Justice, June 9, 2016. https://www.justice.gov/opa/pr/justice-department-and-north-carolina-sue-carolinas-healthcare-system-eliminate-unlawful.

US Department of Transportation. 2017. "DOT Withdraws Two Proposed Rulemakings." News Digest DOT 91-17, Press Office, US Department of Transportation, December 7, 2017. https://www.transportation.gov/briefing-room/dot9117.

US Federal Trade Commission. 2013. "Google Agrees to Change Its Business Practices to Resolve FTC Competition Concerns in the Markets for Devices Like Smart Phones, Games and Tablets, and in Online Search." Press Release, Office of Public Affairs, US Federal Trade Commission, January 3, 2013. https://www.ftc.gov/news-events/press-releases/2013/01/google-agrees-change-its-business-practices-resolve-ftc.

Waldfogel, Joel. 2015. "Digitization and the Quality of New Media Products: The Case of Music." Chap. 14 in *Economic Analysis of the Digital Economy*, edited by Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, 407–42. Chicago: University of Chicago Press.

Weil, David. 2011. "Enforcing Labour Standards in Fissured Workplaces: The US Experience." *Economic and Labour Relations Review* 22(2): 33–54.

Weiss, Leonard, ed. 1990. *Concentration and Price.* Cambridge, MA: MIT Press.

Wollmann, Thomas G. 2019. "Stealth Consolidation: Evidence from an Amendment to the Hart-Scott-Rodino Act." *American Economic Review: Insights* 1(1): 77–94.

Wu, Tim. 2018. *The Curse of Bigness: Antitrust in the New Gilded Age.* New York: Columbia Global Reports.

Yurukoglu, Ali. 2018. "Discussion of 'The Rise of Market Power and the Macroeconomic Implications' by De Loecker and Eeckhout." Presented at the NBER Winter Industrial Organization Meeting, February 9–10, 2018, Stanford, CA.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 69–93

Protecting Competition in the American Economy: Merger Control, Tech Titans, Labor Markets

Carl Shapiro

In the United States, we have a robust set of antitrust laws and antitrust institutions designed to protect and promote competition. These laws date back to the passage of the Sherman Act in 1890, supplemented in 1914 by the Clayton Act and the Federal Trade Commission (FTC) Act. They are enforced by the Antitrust Division of the Department of Justice (DOJ) and by the FTC, together with private antitrust litigation, in which plaintiffs are awarded three times any damages they have suffered. For more than a century, a rich body of case law interpreting these statutes has grown up, heavily influenced by economic research and economic evidence. Indeed, over the twentieth century, the United States led the world in creating and implementing competition policies to control cartels and mergers and to rein in monopoly power.

Yet evidence is mounting that the largest US firms account for a growing share of economic activity, and that profits and price/cost margins at these firms have grown sharply in recent decades. Meanwhile, the economic might of the largest tech firms seems to grow without bound. Have our antitrust laws and institutions failed us?

■ *Carl Shapiro is a Professor at the Haas School of Business, University of California, Berkeley, California. He served as the chief economist at the Antitrust Division of the US Department of Justice during 1995–1996 and 2011–2013 and as a member of the President's Council of Economic Advisers during 2011–2012. He has also served as an economic consultant for the government and for private parties on a variety of antitrust matters over the years. His email address is cshapiro@berkeley.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.69 doi=10.1257/jep.33.3.69

This article makes the case that we need to reinvigorate antitrust enforcement in the United States in three areas. The clearest area where antitrust enforcement has been overly lax is the treatment of mergers. The accumulated evidence indicates that competition would be protected and promoted if the Department of Justice and the Federal Trade Commission were willing and able to block more horizontal mergers. The second area where antitrust enforcement has become inadequate is the treatment of exclusionary conduct by dominant firms. The fundamental problem in this area is that the Supreme Court has, over the past 40 years, dramatically narrowed the reach of the Sherman Act. The third area concerns the market power of employers as *buyers* in labor markets. Historically, antitrust enforcement has largely ignored labor markets. Greater antitrust attention and oversight are warranted, although it is too soon to know whether more robust antitrust enforcement in US labor markets would make a significant difference for the wages earned by employees as a group.

Before discussing antitrust policy in these three areas, it is helpful to lay some groundwork by briefly summarizing some of the evidence that has been accumulating regarding competition and market power in the US economy. Baker (2019) skillfully reviews this evidence in greater detail. Largely on the basis of that evidence, he too advocates for stronger antitrust enforcement.

First, there is clear evidence that corporate profits have risen significantly over the past few decades. The share of US GDP accounted for by corporate profits rose by about half from 1985 to 2016, from around 7.5 to over 11 percent (Shapiro 2018). Barkai (2017) argues that excess profits, meaning the return to capital above the level required to attract investors, have risen sharply as the risk-free rate of return to capital has fallen. Furman and Orszag (2018) report that the premium on the return to all private capital over safe assets has risen from about 200 basis points in 1985 to more than 800 basis points in 2015, and that the return to capital has become much more skewed among US publicly traded nonfinancial firms. The current market capitalizations of the leading US firms further indicate that investors expect these high profits to persist. High and persistent economic profits suggest substantial and durable market power.

Second, there is evidence that price/cost ratios in the United States have risen in recent decades.[1] De Loecker, Eeckhout, and Unger (2018) report a sizable increase in the weighted average ratio of price to marginal cost for publicly traded firms in the United States, from 1.21 in 1980 to 1.61 in 2016, with most of that increase occurring at the top of the distribution. However, measuring price and marginal cost accurately on the basis of public financial accounting data is extremely difficult, especially because large, publicly traded firms produce many products and services.[2] Traina (2018) finds that the ratio of price to marginal cost at nonfinancial, nonutility, publicly traded firms in the United States rose very modestly from 1980

[1] Industrial organization economists have long looked at price/cost margins as indicators of market power. For early reviews and contributions, see Schmalensee (1989), Salinger (1990), and Hall (1988).

[2] The issues surrounding the proper measurement of price/cost ratios are addressed in detail in the articles by Basu; Syverson; and Berry, Gaynor, and Scott Morton in a companion symposium in this issue.

to 2016, from 1.1 to 1.15, if one includes marketing and management expenses when measuring marginal costs. Hall (2018) finds that the weighted average ratio of price to marginal cost increased from 1.12 in 1988 to 1.39 in 2015. The International Monetary Fund (2019), looking across a number of advanced economies, finds that the ratio of price to marginal cost rose by about 8 percent between 2000 and 2016, with that increase concentrated among a small fraction of highly productive and innovative firms. Calligaris, Criscuolo, and Marcolin (2018) obtain similar findings. The antitrust analysis below accepts as a stylized fact that the ratio of price to marginal cost has generally risen over the past 20 to 30 years at the largest, most efficient US firms.

Third, there is convincing evidence that larger, more efficient firms have been growing at the expense of their smaller, less efficient rivals, causing various measures of broad industry concentration in the US economy to increase.[3] Autor et al. (2017a, 185) state, "Our hypothesis is that technology or market conditions—or their interaction—have evolved to increasingly concentrate sales among firms with superior products or higher productivity, thereby enabling the most successful firms to control a larger market share." They call these more efficient firms "superstar firms." Van Reenen (2018) reaches a similar conclusion, based on extensive empirical evidence: "In recent decades the differences between firms in terms of their relative sales, productivity and wages appear to have increased in the US and many other industrialized countries." Bessen (2017) links increases in broad industry concentration to the adoption of information technology. Looking across broad industries, Ganapati (2018) finds a positive correlation between increases in the share of revenue captured by the top four firms and growth in productivity and real output. Crouzet and Eberly (2019) find that the growing role of intangible capital is associated both with the rising share of the largest firms and with productivity gains. None of these conclusions should be surprising, given the extensive literature showing very large and durable differences in efficiency across firms, even in a given industry (Syverson 2011; Van Reenen 2018). The growth of superstar firms may also explain in part the long-term decline in business dynamism (Decker et al. 2016, 2017). The analysis below accepts as a stylized fact that the advantages enjoyed by efficient market leaders over followers and entrants have grown considerably over time.

Fourth, labor's share of GDP has significantly declined since the 1980s. Autor et al. (2017a, b) show that this decline is due to the reallocation of activity toward firms with low and declining labor shares, namely the superstar firms that have

[3] This Darwinian mechanism does not appear to be operating as strongly in Europe as in the United States. Europe has not experienced a similar increase in the share of economic activity accounted for by the largest firms. Gutiérrez and Philippon (2018) report that from 1995 to 2015, the weighted average Herfindahl–Hirschman index (HHI) for the United States measured at the broad sector level rose from about 500 to about 650, while the comparable HHI in each of ten large EU countries fell from about 800 to about 600. (They find lower HHIs in the EU overall than in individual member countries.) Likewise, Valletti (2018) finds no increase in concentration in EU countries from 2010 to 2015 at the broad sector level. Note that these measures of concentration are uninformative regarding market power because these broad sectors bear no relation to relevant product markets.

gained market share in recent years. The declining labor share also naturally raises the question of whether employers have growing market power in labor markets, especially given the declining role of unions, driving down wages and exacerbating inequality in the distribution of income and wealth.

None of these trends necessarily indicates that US antitrust policy has been deficient. Indeed, the fact that price/cost margins have risen in many high-income countries suggests that growing economies of scale and globalization are the cause, not domestic policy changes. Nonetheless, these trends compel us to take a closer look at recent antitrust enforcement in the United States, to ask whether stronger enforcement is now needed, and to see how that can be accomplished.

This article does not address one core aspect of antitrust enforcement: the prohibition on cartels and price-fixing. The Department of Justice regularly brings criminal charges against individuals who engage in price-fixing. Companies found to have participated in cartels are subject to fines assessed by the government plus liability for treble damages to customers who were overcharged. Over the past ten years, the DOJ's Antitrust Division has assessed roughly $10 billion in criminal fines and penalties (Department of Justice 2019) and has thrown quite a few executives in jail. This experience shows that antitrust vigilance regarding cartels is vital to a competitive economy. Cartels continue to form and persist in many industries. Levenstein and Suslow (2006, 2011) find that the average duration of cartels is about eight years. Improved detection of active cartels would do much to promote competition.

Merger Control

Under the Clayton Act, mergers that "may substantially lessen competition" are illegal in the United States. Merger control works together with the criminal prohibition on cartels to protect competition. Without merger control, rivals could achieve collusive outcomes by merging.

In 2017, about 15,000 merger and acquisition deals were announced in the United States, representing about $2 trillion in total value. Merger control policy greatly affects the set of deals that are proposed as well as which deals obtain antitrust clearance and are consummated.

The Department of Justice and the Federal Trade Commission have the authority to investigate and challenge proposed mergers *before* the merging parties are permitted to consummate their merger. If the DOJ or FTC can convince a federal judge that a merger is anticompetitive, that merger is blocked. Economic analysis plays a central role in this process. The DOJ and the FTC publish Horizontal Merger Guidelines that explain to the business community and the courts how they analyze horizontal mergers (Department of Justice and Federal Trade Commission 2010).

In 2017, 2,052 proposed transactions were reported to the antitrust authorities, of which 51 received an in-depth investigation and 21 were subject to an enforcement action (Federal Trade Commission and Department of Justice 2018).

Challenged deals often go forward after the merging parties agree to a remedy, usually an asset divestiture designed to preserve competition. Some deals are abandoned in the face of an antitrust challenge, such as the proposed merger between AT&T and T-Mobile in 2011 and the one involving Halliburton and Baker Hughes in 2016. Very few deals are litigated in court.

The fundamental challenge for merger control is that it is a *predictive exercise*: if one is seeking to identify the subset of proposed mergers that "may substantially lessen competition," one must assess the likely competitive effects of a proposed merger *before* it is consummated.

The Gradual Weakening of Merger Control

Fifty years ago, predictions of merger effects were based largely on market shares, and merger control was very strict. In 1963, relying heavily on the economics literature, the Supreme Court held that any merger producing a firm that controls an "undue percentage share" of the market and that "results in a significant increase in the concentration of firms in that market" is "inherently likely to lessen competition substantially" (*United States v. Philadelphia National Bank*, 374 US 321 [1963]). Such a merger would be presumed to be illegal "in the absence of evidence clearly showing that the merger is not likely to have such anticompetitive effects." This established a "structural presumption" against mergers based on market concentration.

As a prominent—some would say infamous—example of the structural presumption in action, the Supreme Court upheld the Department of Justice challenge to a merger between two grocery chains with a combined market share of 7.5 percent in the retail grocery market in the Los Angeles area (*United States v. Von's Grocery*, 384 US 270 [1966]). One year later, the Court ruled out efficiencies as a reason for allowing a merger, stating that "[p]ossible economies cannot be used as a defense to illegality" (*FTC v. Procter & Gamble*, 386 US 568 [1967]). See Hovenkamp and Shapiro (2018) and Werden (2018) for discussions of merger law in the 1960s.

The merger enforcement policies of the Department of Justice during the 1960s and 1970s reflected these Supreme Court rulings. When the DOJ published its first Merger Guidelines in 1968, they focused very heavily on the market shares of the merging firms. They stated, for example, that the DOJ "will ordinarily challenge" a merger between two firms with 5 percent market share each, or between a firm with a 20 percent market share and a firm with a 2 percent market share.

Since 1968, merger enforcement has evolved significantly along two distinct dimensions. First, the level of market concentration required to trigger the structural presumption has risen. Few if any antitrust economists today would favor applying the low thresholds found in the 1968 Merger Guidelines, given what we now know about the effects of horizontal mergers. Second, merger analysis now puts less weight on market shares and more weight on other evidence to predict the competitive effects of a merger. This shift reflects the accumulation of experience at the Department of Justice and the Federal Trade Commission along with the recognition that each industry has unique features. In Shapiro (2010), I explain this shift and how it took place.

Remarkably, the treatment of proposed mergers under US antitrust law has become much more lenient without Congress changing the substantive standard to be used for merger review and without any updated guidance from the Supreme Court, which has not heard a merger case since 1974. Rather, these changes have resulted from a dynamic involving the lower courts, the antitrust agencies, antitrust lawyers, and economists.

That process began in earnest in 1982, when the Department of Justice dramatically revised the Merger Guidelines, giving less weight to market shares and raising the threshold levels of concentration that would trigger the structural presumption. Within a decade, the courts followed the DOJ's lead. In 1990, the most influential lower court, the DC Circuit, rather brazenly departed from the precedent established by the Supreme Court in the 1960s, stating that "[e]vidence of market concentration simply provides a convenient starting point for a broader inquiry into future competitiveness" (*United States v. Baker Hughes*, 908 F.2d 981, at 984). This ruling weakened the structural presumption, making it harder for the DOJ and Federal Trade Commission to block mergers.

The Department of Justice, joined by the Federal Trade Commission, further revised its Horizontal Merger Guidelines in 1992, 1997, and 2010.[4] With each revision, less weight was given to market shares and greater weight was attached to more direct evidence about how competition has taken place in the industry and how the merger would likely alter that competition. The 1992 guidelines introduced "unilateral effects" into the analysis, shifting attention to the loss of direct competition between the merging firms and away from the overall structure of the market. The 2010 guidelines introduced the concepts of upward pricing pressure, merger simulation, and bidding and auction models into the analysis of unilateral effects. They also addressed nonprice dimensions of merger analysis, including product variety and innovation. In Shapiro (2010), I describe how the 2010 guidelines built on decades of experience with the economic analysis of the competitive effects of mergers, and in Shapiro (2012), I emphasize the importance of competition in spurring innovation.

In principle, the antitrust agencies can more accurately distinguish procompetitive horizontal mergers from anticompetitive ones by undertaking detailed analyses of how a proposed merger is likely to alter competition. However, as the Department of Justice and the Federal Trade Commission have put more weight on direct evidence of competitive effects and less weight on market shares, a gap has opened up between how they evaluate mergers and how the courts do so. Economists at the antitrust agencies engage in sophisticated analysis, interacting with economists hired by the merging parties. But if a DOJ or FTC merger challenge is litigated, a generalist judge may then be faced with conflicting expert testimony that is difficult to decipher.

[4] I led the team at the Department of Justice that drafted the 2010 Horizontal Merger Guidelines. Joseph Farrell, also a professor at the University of California, Berkeley, led the team at the Federal Trade Commission.

The result is that the antitrust authorities still rely heavily on market definition, market shares, and the structural presumption to make their case in court, even when their enforcement decisions are based on other economic evidence, such as bidding data, upward pricing pressure, or merger simulations. But the structural presumption has become weaker as the lower courts place less weight on market concentration and increasingly look for direct evidence of the likely effects of proposed mergers (Hovenkamp and Shapiro 2018). The net result is that the antitrust authorities have found it harder to prevail in court, causing them to be more cautious in the mergers they challenge. Merging firms understand this, and the mix of proposed mergers has adjusted accordingly. This is now highly problematic, given the mounting evidence cited above about rising profits, widening price/cost margins, and the rise of superstar firms.

The Need for Stronger Merger Enforcement in an Economy with Superstar Firms

Accumulating evidence in two broad categories points to the need for more stringent horizontal merger enforcement policy in the United States: (1) evidence showing that the largest and most successful US firms have increasing market power and (2) evidence from merger retrospectives.[5]

The evidence cited above shows that superstar firms are highly profitable owing to durable competitive advantages they enjoy over their smaller rivals and over entrants, which cannot easily or quickly replicate their assets and capabilities. These are precisely the conditions under which mergers involving successful established firms are most likely to lessen competition and harm customers. We also know that higher price/cost margins cause the unilateral price effects from horizontal mergers to be more harmful to customers. Likewise, research and development competition spurs innovation if future sales are contestable (Shapiro 2012; Federico, Scott Morton, and Shapiro forthcoming), so a merger between two firms investing to develop competing products is likely to slow down innovation.

The case for stronger merger enforcement is perfectly consistent with a conclusion that the rise of superstar firms has largely resulted from the normal competitive process in the presence of growing economies of scale, increased globalization, and dramatic improvements in information technology, combined with large and persistent differences across firms in their ability to adopt new technologies and adjust to changing market conditions. After all, we expect a healthy competitive process to result in the most efficient firms gaining market share through internal growth and earning above-normal profits. As emphasized by Valletti and Zenger (2019), merger controls preserve that competition.

Contrary to many popular views, the case for stronger merger enforcement does *not* rest on evidence showing that various broad US industries have become more concentrated over time. Measures of industry concentration based on data

[5] Azar, Schmalz, and Tecu (2018) and Schmalz (2018) argue that growing common ownership of rivals by financial firms has weakened rivalry in many oligopolistic markets. If this claim finds additional support in future research, it would provide an additional basis for a more stringent merger control policy.

from the US Economic Census are simply not very informative for merger analysis because these data are available only at an aggregated level. The modest increases in concentration observed when using these data confirm that the largest firms are responsible for a greater portion of economic activity in many industries, but they tell us very little about concentration in properly defined relevant antitrust markets (Shapiro 2018).

As one important illustration of this point, a great many real-world markets—such as hospital services, supermarkets, and automobile dealers—are local. Reported changes in *national* concentration can vastly overstate or understate changes in concentration in local markets. A local merger creating a local monopoly would not even show up in the national measures. Conversely, if large regional firms are growing and entering many local markets, their entry would cause concentration in those local markets to *fall*. Rossi-Hansberg, Sarte, and Trachter (2018) report that "the *positive* trend observed in national product-market concentration between 1990 and 2014 becomes a *negative* trend when we focus on measures of local concentration."

As another illustration, North American Industry Classification System (NAICS) 3254 is "Pharmaceutical and Medicine Manufacturing," which encompasses a very large number of drugs that are not substitutes for each other. Reported changes in concentration in this four-digit industry tell us little or nothing about concentration in the supply of drugs to treat specific ailments. One simply cannot detect overall trends in concentration in properly defined relevant markets using data from the Economic Census.

Furthermore, it is important to remember that an increase in *concentration* in a properly defined relevant market does not prove that *competition* in that market has declined. Consider an unconcentrated market in which a few of the many suppliers become more efficient and gain share by offering lower prices and improved products, causing concentration to rise. That increase in concentration clearly goes hand in hand with customer benefits and reflects the competitive process at work, not a decline in competition. Industrial organization economists have understood this fundamental point for at least 50 years, and probably much longer (as one leading example, see Demsetz 1973). To properly interpret an observed increase in market concentration, one must understand what *caused* concentration to rise. Rising concentration resulting from increased efficiency by one or a few firms reflects the competitive process at work; rising concentration caused by mergers may well reflect a decline in competition. Distinguishing one fact pattern from the other requires looking at properly defined individual markets and examining what has actually happened over time in those markets.

In any event, regardless of whether one loves or loathes the rise of superstar firms, the implications for merger analysis are unambiguous: proposed horizontal mergers involving highly successful firms should be greeted with considerable skepticism.

The second broad category of evidence supporting more robust horizontal merger enforcement comes from merger retrospectives: studies of the economic effects of consummated mergers. The most convincing studies identify a control product or geography and use a difference-in-difference methodology to isolate the merger's effect on prices. For example, Ashenfelter and Hosken (2010) look at five mergers this way, four of which resulted in price increases for consumers. Likewise, Whirlpool's acquisition of Maytag led to relatively large price increases for clothes dryers (Ashenfelter, Hosken, and Weinberg 2013). Blonigen and Pierce (2016) find that mergers in the manufacturing sector have generally been associated with increases in markups but not with increases in productivity.

Mergers in the health-care sector have been especially harmful to competition. Gaynor and Town (2012) report that hospital mergers in concentrated markets typically lead to price increases of at least 20 percent compared with control hospitals. This literature also shows that hospital competition improves the quality of care. Gaynor (2018) presents evidence that "consolidation between close competitors leads to substantial price increases for hospitals, insurers, and physicians, without offsetting gains in improved quality or enhanced efficiency." The Center for American Progress (Gee and Gurwitz 2018) assembles a range of evidence also supporting this conclusion.

Kwoka (2014) provides the most comprehensive review of merger retrospectives. Most mergers in these studies are older ones that took place in a few industries where data were available: banking, hospitals, airlines, petroleum, and journal publishing. Kwoka reports that "most studied mergers result in competitive harm, usually in the form of higher prices" (158). Ashenfelter, Hosken, and Weinberg (2014) review these studies, stating that "[t]he empirical evidence that mergers can cause economically significant increases in price is overwhelming" (S78).

While the overall body of evidence from merger retrospectives, standing alone, does not allow us to predict with confidence the effects of any given merger, it does indicate that merger enforcement has been too lax over the past 25 years.

How to Reinvigorate Merger Enforcement

Merger enforcement can be strengthened in a number of ways, if the Department of Justice and the Federal Trade Commission choose to move in that direction and if the courts cooperate.

First, the structural presumption against mergers that increase concentration in a properly defined relevant market could be strengthened. For example, after the government has defined a relevant market based on substantial evidence, the merging parties could be required to present clear and convincing evidence to contest that definition of the market. Similarly, once the government has established the structural presumption, the merging parties could be required to present clear and convincing evidence to rebut the presumption. Claims by the merging parties that growth by small rivals or entry of new firms into the market will quickly and effectively restore any competition lost due to the merger could be greeted with greater skepticism.

Second, more weight could be placed on evidence that the merging parties are significant direct competitors, without the necessity of defining a relevant market and measuring market shares in that market, and without requiring the government to quantify the harm to customers that the merger will cause. For example, strong evidence that customers have often obtained lower prices as a result of direct competition between the two merging firms could, at least in some cases, be regarded as sufficient for the government to meet its initial burden of showing that the merger "may substantially lessen competition."

Third, the agencies and the courts could express greater wariness when a dominant incumbent firm seeks to acquire a firm operating in an adjacent market, especially if the target firm is well positioned to challenge the incumbent's position in the foreseeable future. In the language of antitrust law, this would involve lowering the evidentiary requirements necessary for the government to prevail in a merger case based on a loss of "potential competition." For example, the government could meet its initial burden by showing that the target firm is reasonably likely to become a rival to the acquiring firm in the foreseeable future, even if the target firm has not yet made specific plans to do so. This change would reduce the ability of powerful firms to acquire potential rivals before they mature into actual rivals, without stopping them from making acquisitions to improve their offerings or to challenge other firms with entrenched positions.

This change would be especially consequential as applied to dominant firms in the tech sector. Under this standard, Facebook's acquisitions of Instagram and WhatsApp might well have been blocked, if these firms were seen as well placed to mature into rivals to Facebook as social media platforms, and Google's earlier acquisitions of YouTube and DoubleClick would at least have warranted greater scrutiny. But it seems unlikely that Amazon's acquisition of Whole Foods or Google's acquisition of Nest would have raised serious issues even under this stricter standard. More acquisitions by the tech titans involving important inputs or complements could be challenged, but it is unclear how the courts would respond to cases involving such vertical or complementary mergers.

Fourth, the courts could insist that any "fixes" to proposed mergers result in a market structure that preserves competition. This would involve skepticism about divestitures designed to obtain antitrust approval that do not make good business sense, and placing little or no weight on behavioral commitments by the merged entity, such as a commitment not to raise price.

Lastly, the Department of Justice and the Federal Trade Commission could be given more resources to investigate and challenge mergers. With more resources, the antitrust agencies could look more closely at more suspect proposed mergers. They also could investigate more *consummated* mergers to see whether they have harmed competition or are likely to do so, including mergers that were below the size-of-transaction reporting threshold, which is $90 million in 2019. Cunningham, Ederer, and Ma (2018) find that acquired pharmaceutical projects are less likely to be developed when they overlap with the product portfolio of the acquiring firm; these "killer acquisitions" disproportionately occur just below the reporting threshold. Wollmann (2019) also provides worrisome evidence about mergers taking place just below the threshold.

As a practical matter, the case law relating to mergers evolves very slowly, with substantial lags following advances in economic learning and then changes in Department of Justice and Federal Trade Commission merger enforcement policies. Whether the current judiciary has the appetite to support stronger merger enforcement remains to be seen. If not, or if that route is too slow, Congress would need to pass new legislation codifying changes such as the ones suggested above—a heavy lift to be sure.

Antitrust and the Tech Titans

The most talked-about antitrust question of the day is whether and how antitrust should act to limit the economic power of the largest tech firms, often identified as Amazon, Apple, Facebook, and Google. Should they be broken up? Forced to modify their business practices and pay fines for their past sins? Watched carefully? Left alone and applauded?

A first step toward answering these questions is to recognize that the goal of antitrust policy is to protect and promote competition. Antitrust is not designed or equipped to deal with many of the major social and political problems associated with the tech titans, including threats to consumer privacy and data security, or with the spread of hateful speech and fake news. Indeed, it is not clear that more competition would provide consumers with greater privacy or would better combat information disorder; unregulated competition might instead trigger a race to the bottom, and many smaller firms might be harder to regulate than a few large ones. Addressing these major problems requires sector-specific regulation, which is beyond the scope of this article.

Three important and insightful reports have recently been released addressing antitrust in the digital economy: one by the United Kingdom (Furman et al. 2019), one by the European Commission (Crémer, de Montjoye, and Schweitzer 2019), and one by the Stigler Center at the University of Chicago (Scott Morton et al. 2019). All three reports conclude that antitrust can and should do more to promote competition in the digital era, while staying true to its focus on competition issues. All three reports call for regulation to address other public policy issues relating to digital platforms.

Within the realm of antitrust, it is important to understand that under the Sherman Act, it is not illegal for a company to have a monopoly, so long as that position was achieved by offering customers attractive products and services. There is a broad consensus behind this approach, because it would be illogical to urge companies to compete and then tell them they have broken the law when they do so successfully. Judge Learned Hand famously captured this idea in *United States v. Aluminum Company of America* (148 F.2d 416, Second Circuit [1945]):

> A single producer may be the survivor out of a group of active competitors, merely by virtue of his superior skill, foresight and industry. In such cases a

> strong argument can be made that, although the result may expose the public to the evils of monopoly, the Act does not mean to condemn the resultant of those very forces which it is its prime object to foster: *finis opus coronat*. The successful competitor, having been urged to compete, must not be turned upon when he wins.

Following this core principle, the basic antitrust question for each tech titan is whether that company has engaged in practices that go beyond competition on the merits and are likely to (1) exclude its rivals and fortify its market position or (2) extend its power to adjacent markets. If so, a remedy is needed to restore competition. A behavioral remedy imposes limits and obligations on the company's conduct; this was the outcome in the Microsoft case 20 years ago. A structural remedy breaks up the company; this was the result in the AT&T case in the 1980s. Talk of breaking up the tech titans without reference to a specific antitrust violation is putting a very large cart before the horse.[6]

The Shrinking Scope of the Sherman Act

The portion of the Sherman Act dealing with monopolies is remarkably broad—and vague. Section 2 states, "Every person who shall monopolize, or attempt to monopolize . . . any part of the trade or commerce among the several States . . . shall be deemed guilty of a felony."

From the outset, it was clear that the courts would play a major role in interpreting the broad language of the Sherman Act. As discussed in the companion piece in this symposium by Naomi Lamoreaux, the role of antitrust in the American economy has waxed and waned depending on judicial rulings, which have evolved in response to economic learning and changing market conditions as well as political forces and the makeup of the Supreme Court.

For many years, the Supreme Court recognized the expansive nature of the antitrust statutes. In 1958, the Court stated, "The Sherman Act was designed to be a comprehensive charter of economic liberty aimed at preserving free and unfettered competition as the rule of trade" (*Northern Pacific Railway v. United States*, 356 US 1 [1958], at 4). In 1972, the Court stated, "Antitrust laws in general, and the Sherman Act in particular, are the Magna Carta of free enterprise. They are as important to the preservation of economic freedom and our free-enterprise system as the Bill of Rights is to the protection of our fundamental personal freedoms" (*United States v. Topco*, 405 US 596 [1972], at 610).

The high-water mark for antitrust in the United States was reached during the 1960s. Since then, the Supreme Court has substantially narrowed the scope of the antitrust laws. This narrowing took place along multiple dimensions.

[6] As discussed above, a breakup to unwind a prior anticompetitive acquisition could well be a suitable remedy and would not be novel as a legal matter. This section considers exclusionary conduct by dominant firms, not mergers.

A number of business practices that previously were treated as automatically or per se anticompetitive are now evaluated on a case-by-case "Rule of Reason" basis. For example, it used to be per se illegal for a manufacturer to assign territories to its distributors and to prevent one distributor from selling outside its assigned territory. But in 1977, the Supreme Court ruled that this practice would in the future be evaluated using the Rule of Reason (*Continental TV v. GTE Sylvania,* 433 US 36 [1977], overruling *United States v. Arnold Schwinn and Company,* 388 US 365 [1967]). Likewise, retail price maintenance—when a manufacturer prohibits a retailer from selling its products below a specified price—used to be per se illegal. In 2007, the Court ruled that it would be evaluated using the Rule of Reason (*Leegin Creative Leather Products v. PSKS,* 551 US 877 [2007], overruling *Dr. Miles Medical Company v. John D. Park & Sons Company,* 220 US 373 [1911]).

The Court also erected obstacles to antitrust plaintiffs in situations when the Rule of Reason is applied. For example, in 1993 the Court ruled that a plaintiff in a predatory pricing case had to show that the monopolist was selling below cost *and* that the monopolist would be able to recoup the losses resulting from this below-cost pricing by charging higher prices in the future (*Brooke Group v. Brown & Williamson Tobacco Corporation,* 509 US 209 [1993]). Under this standard, a predatory pricing case by the Department of Justice against American Airlines failed (*United States v. AMR Corporation,* 335 F.3d 1109, Tenth Circuit [2003]). Similarly, in 2004, the Court significantly narrowed the set of circumstances in which an antitrust plaintiff could win on the basis of a monopolist's refusal to sell an essential input to a rival (*Verizon Communications v. Law Offices of Curtis V. Trinko,* 540 US 398 [2004], narrowing *Aspen Skiing v. Aspen Highlands Skiing,* 472 US 585 [1985]).[7]

Collectively, these and other cases represent a significant backing away from antitrust by the Supreme Court. Baker (2015) argues forcefully that many of these judicial decisions were based on erroneous assumptions, including these: that markets self-correct through entry, that oligopolists compete and cartels are unstable, and that business practices prevalent in competitive markets cannot harm competition. This history suggests that it will be challenging for the government to bring a successful Sherman Act case against the tech titans. But ultimately that will depend on the specific facts of the case. A strong case can withstand these headwinds.

The Microsoft case (*United States v. Microsoft,* 253 F.3d 34, DC Circuit [1998]) provides the best guide to what constitutes monopolization in a high-tech setting. This case should be encouraging for those in favor of antitrust action against the tech titans. Microsoft was found to have monopolized the market for Intel-compatible operating systems for personal computers, on the basis of conduct that excluded the Netscape browser and Java software, which together might have facilitated entry

[7] Based on *Trinko,* the Court subsequently limited the circumstances under which a plaintiff can bring an antitrust case based on a price squeeze (see *Pacific Bell Telephone Company v. LinkLine Communications,* 555 US 438 [2009]).

and thus eroded the monopoly power of Microsoft Windows. For an overview of the arguments, see the three-paper symposium in the Spring 2001 issue of this journal.

Applying Antitrust Principles to the Tech Titans

We are now ready to look more closely at Google, Amazon, Facebook, and Apple. These four "GAFA" firms have received by far the most antitrust attention of late as they have become economically and socially important. However, they are not the only firms with powerful positions in the information economy. One could also look for exclusionary conduct by other large high-tech firms, including Microsoft, Oracle, IBM, Salesforce, Adobe, and Uber. Expanding to telecom and media, the list could include AT&T, Verizon, Comcast, Netflix, and Disney. One also could conduct a similar exercise in other industries, such as pharmaceuticals, semiconductors, or airlines.

The point of this discussion is not to offer a view on whether any of these four GAFA firms have violated US antitrust law, but rather to show how antitrust principles can be applied to these tech titans. The analysis is illustrative and is not intended to capture all plausible antitrust cases that might be brought against these firms. Each of these companies is assumed to have sufficient market power to be subject to Section 2 of the Sherman Act. However, is important to remember that these firms' products and services, and their business models, are very different from one another, so it makes no sense to lump them together. Any antitrust analysis must be done company by company, based on that company's practices. Here we briefly consider the potential application of three antitrust doctrines to the tech titans: predatory pricing, exclusion of nascent threats, and extension of market power into adjacent markets.

A monopolist that engages in *predatory pricing*—that is, pricing below cost to drive rivals out of the market so it can then recoup those losses by raising prices to monopoly levels—can be guilty of monopolization. One can ask whether Amazon's (or Uber's) conduct falls into this category, as some have alleged.

Amazon's core online retailing business has clearly generated enormous benefits to consumers by offering low prices, a huge variety of products, and reliable and speedy delivery. Amazon has been a highly successful company, putting great competitive pressure on other retailers. Under Supreme Court precedent, a predatory pricing case against Amazon would fail unless Amazon were shown to have priced below cost. Furthermore, showing the prospect of future harm to consumers would also be a necessary element of any case. Requiring plaintiffs in predatory pricing cases to show some prospect of harm to customers, not just harm to competitors, is critical to the coherence of antitrust policy, to avoid chilling legitimate price competition.[8] Showing such harm to consumers from Amazon's aggressive pricing

[8] In an earlier era, the grocery chain A&P offered consumers a wide range of products at low prices, putting pressure on many smaller grocery stores. A&P was successfully prosecuted by the US Department of Justice, an action now widely seen as misguided (Muris and Neuchterlein 2018). Later, Walmart was the innovator in retailing, putting great competitive pressure on smaller retailers around the country.

and growth strategy would appear to be difficult. Hemphill and Weiser (2018) offer a road map to bringing and deciding predatory pricing cases.

A tech titan also could be challenged for *exclusion of nascent threats*, which means using its dominant position to exclude products or services that it fears may grow to threaten its core business. This was the basic economic logic of the case against Microsoft 20 years ago. The Netscape browser and Java were the potential threats to Microsoft Windows. The *Microsoft* case established antitrust liability for a dominant firm that excludes rivals, even if the threats they pose are "nascent." But the reach of the *Microsoft* case is unclear, since the Supreme Court subsequently ruled in the *Trinko* case that a dominant firm normally has no duty to deal with its rivals. If the Supreme Court applies *Trinko* broadly to the tech titans, then separate regulation might be needed to impose on the tech titans mandated interconnection or data sharing with rivals.

One way a dominant firm can exclude rivals is by refusing to sell its product to customers who also purchase from its rivals. In 1951, the Supreme Court ruled that such "exclusive dealing" violated Section 2 of the Sherman Act. That case involved a dominant local newspaper that refused to accept advertisements from those who also placed advertisements on the local radio station (*Lorain Journal v. United States*, 342 US 143 [1951]). A tech titan putting up obstacles to customers seeking to also use rival products could easily face liability under this precedent. As a recent example of exclusionary conduct, Facebook blocked Vine, a video sharing app launched by Twitter in January 2013, from accessing Facebook user data (O'Sullivan and Gold 2018). This prevented Facebook users from inviting their Facebook friends to join Vine. Facebook was applying its policy of restricting access to apps that replicated Facebook's core functionality. In response to the Vine episode becoming public, Facebook stated that it was dropping this policy (Facebook 2018), which appears difficult to defend. Twitter discontinued the Vine mobile app in October 2016.

A third category of candidate antitrust cases against the tech titans involves allegations that a firm is *abusing its dominant position to expand into adjacent markets.* Several tech titans already face cases that focus on whether they have favored their own service over a rival service in an adjacent market.

The European Commission is investigating whether Amazon is using data collected about third-party sellers on its platform to guide its own product offerings in competition against those third-party sellers. Such conduct could be problematic, but note that when one firm simply imitates its rivals, that is normally an important channel for the diffusion of new ideas, so long as the imitation does not involve any breach of contract or infringement of intellectual property rights. Looking more broadly, Amazon could be accused of favoring its own products over those of third parties selling on its platform, by giving its own products preferred placement on Amazon's website or by charging third parties excessive rates to be fulfilled or sold through Amazon. This last type of case would be very difficult in the United States under Supreme Court precedent.

Apple has been accused of discriminating against rivals who rely on the Apple platform to reach consumers. In March 2019, the music streaming service Spotify filed an antitrust complaint at the European Commission against Apple (Ek 2019). Spotify objected to the 30 percent fee that Apple charges on certain purchases made through Apple's payment system and claimed that Apple had locked Spotify out of Apple Watch. Spotify asserted that it should receive the same treatment at the Apple App Store given to Apple's competing service, Apple Music. In response, Apple claimed that it had worked closely with Spotify for years and that Spotify was not willing to abide by the same rules that apply to all apps on the App Store, which Apple regards as necessary for the operation and security of the App Store. Apple further claimed that Apple approved Spotify for the Apple Watch and that Spotify has been the leading app in the Watch Music category. The European Commission is opening an investigation in response to Spotify's complaint.

The Spotify complaint illustrates the tensions that arise when the company controlling a platform also offers its own services on that platform. Indeed, the boundary between the "platform" and services running on that platform can be fuzzy and can change over time. Similar issues will surely arise for other applications that rely on Apple's App Store to reach customers. For example, Apple recently removed several parental control apps from the App Store. These apps provide alternatives to Apple's own screen-time control tools. Apple explained that it took this action to protect users' privacy and security, but an antitrust complaint here would not be shocking (Apple 2019).

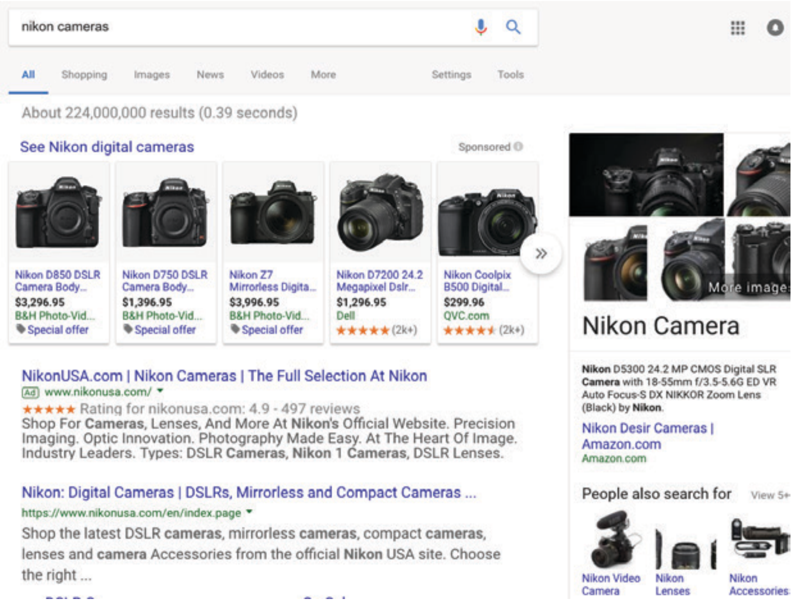
We know a lot more about what a case of this type might look like against Google, because the European Commission issued an antitrust decision in June 2017 against Google involving Google Shopping, including a €2.42 billion ($2.7 billion) fine.[9] Google displays advertisements when users enter queries into the Google search engine that relate to commercial products. For example, Figure 1 shows what one sees on a desktop computer if one searches for "Nikon Cameras" on Google.

All of the images displayed in Figure 1 are *sponsored search results*; that is, they are advertisements paid for by online merchants. Google calls these "Product Listing Ads." A user who clicks on one of these ads is directed to the website of the online merchant sponsoring that ad, and that sponsor pays a fee to Google. The first link to NikonUSA.com is also an advertisement, in text form. The second link to NikonUSA.com is a *generic search result* generated by Google's algorithm, not an advertisement. More generic search results, not shown in Figure 1, follow.

Many press reports have left the impression that the European Commission case was about Google biasing its search algorithm by demoting its rivals, but that is not correct. The European Commission fact sheet states, "The Commission Decision does not object to the design of Google's generic search algorithms or to demotions as such, nor to the way that Google displays or organizes its search results pages (e.g., the display of a box with comparison shopping results displayed prominently in a rich, attractive format)" (European Commission 2017). Instead, the European Commission "objects to the fact that Google has leveraged its

[9] I served as an economic expert for Google in the Google Shopping case. The views expressed here are my own.

Figure 1
Result of a Search for “Nikon Cameras” on Google



About 224,000,000 results (0.39 seconds)

See Nikon digital cameras

Nikon D850 DSLR Camera Body...
$3,296.95
B&H Photo-Vid...
Special offer

Nikon D750 DSLR Camera Body...
$1,396.95
B&H Photo-Vid...
Special offer

Nikon Z7 Mirrorless Digita...
$3,996.95
B&H Photo-Vid...
Special offer

Nikon D7200 24.2 Megapixel Dslr...
$1,296.95
Dell
★★★★★ (2k+)

Nikon Coolpix B500 Digital...
$299.96
QVC.com
★★★★★ (2k+)

NikonUSA.com | Nikon Cameras | The Full Selection At Nikon
Ad www.nikonusa.com/
★★★★★ Rating for nikonusa.com: 4.9 - 497 reviews
Shop For **Cameras**, Lenses, And More At **Nikon's** Official Website. Precision Imaging. Optic Innovation. Photography Made Easy. At The Heart Of Image. Industry Leaders. Types: DSLR **Cameras**, **Nikon** 1 **Cameras**, DSLR Lenses.

Nikon: Digital Cameras | DSLRs, Mirrorless and Compact Cameras ...
https://www.nikonusa.com/en/index.page
Shop the latest DSLR **cameras**, mirrorless **cameras**, compact **cameras**, lenses and **camera** Accessories from the official **Nikon** USA site. Choose the right ...

Nikon Camera

Nikon D5300 24.2 MP CMOS Digital SLR **Camera** with 18-55mm f/3.5-5.6G ED VR Auto Focus-S DX NIKKOR Zoom Lens (Black) by **Nikon**.
Nikon Desir Cameras | Amazon.com
Amazon.com

People also search for

Source: Unaltered screen capture of a Google search for “Nikon Cameras” performed by the author in November 2018.
Notes: All of the images in the main column are *sponsored search results*; that is, they are advertisements paid for by online merchants. A user who clicks on one of these ads is directed to the website of the online merchant sponsoring that ad, and that sponsor pays a fee to Google. The first link to NikonUSA.com is also an advertisement, in text form. The second link to NikonUSA.com is a *generic search result* generated by Google’s algorithm, not an advertisement. Additional generic search results, not shown in the figure, follow.

market dominance in general internet search into a separate market, comparison shopping. Google abused its market dominance as a search engine to promote its own comparison shopping service in search results, whilst demoting those of rivals.”

According to the European Commission, Google did this by displaying Product Listing Ads, such as those shown in Figure 1. This is a peculiar claim, because those Product Listing Ads are very much like the text ads that Google has shown for years, and the European Commission does not object to ads that use text rather than images. Plus, as the European Commission recognizes, there is nothing wrong from a competitive perspective when a content provider earns revenue by selling advertisements. The newspaper and radio industries have done that for a very long time. Furthermore, it is not apparent how the Product Listing Ads “promote” Google’s comparison shopping service, since a user who clicks on one of those ads is directed to the merchant’s website, not to the stand-alone Google Shopping site.

Lessons

We can distill several lessons regarding the role and limits of antitrust in controlling the tech titans.

First, many of the deepest concerns about the tech titans, including privacy, data security, and information disorder, do not directly involve competition issues. Sector-specific regulation is overdue and badly needed to address these problems. Antitrust cannot solve all manner of economic and social problems and should not be expected to do so.

Second, those who would like to see the tech titans broken up on antitrust grounds are likely to be disappointed, since antitrust does not condemn monopoly as such. Antitrust liability requires that a dominant firm abuse its power in some way, such as by excluding rivals. When liability is found, a suitable remedy is designed to restore the competition lost due to the illegal acts.

Third, establishing that a dominant firm has abused its position has become much harder in the United States over the past 40 years, under a series of decisions by the Supreme Court. This has left antitrust enforcement in the United States notably weaker than in the European Union. That is unlikely to change much in the near future, so look to Brussels for much of the action.

Fourth, many of the antitrust cases against the tech titans in the years ahead will most likely involve allegations that these firms have used their "platforms" to favor their own services over rival services. These cases will be complex, and they will be risky for plaintiffs in the United States given Sherman Act case law and the current makeup of the Supreme Court.[10]

Despite these obstacles, ongoing antitrust oversight and vigilant antitrust enforcement toward the tech titans is critical and can make a real difference, mostly through deterrence rather than litigation. Antitrust can and should prevent the tech titans from entrenching their economic power by engaging in exclusionary conduct that weakens the competitive pressures they face from rivals offering new and disruptive products and services. At the same time, antitrust should take care not to discourage the tech titans from competing with each other, as Microsoft has done with Bing against Google and as Google has done with Android against Apple iOS.

Antitrust in Labor Markets

Antitrust enforcement in the United States has largely focused on the market power of *sellers*. However, the Sherman Act applies with equal force to the market

[10] Adding to concerns, the Supreme Court recently issued a worrisome and poorly reasoned decision involving payment systems that could greatly complicate any case brought in a "two-sided market" (*Ohio v. American Express*, 138 S. Ct. 2274 [2018]). Justice Breyer's dissent in this case eviscerates the majority opinion. While it is not yet clear how broadly the lower courts will read the *American Express* decision, all four of the tech titans arguably operate in "two-sided markets." Apple connects users with application developers, Facebook and Google connect advertisers with users, and Amazon connects manufacturers and merchants with consumers.

power of *buyers*. When Weyerhaeuser was sued as a buyer for bidding up the price of sawlogs in the Pacific Northwest to a level that prevented rival sawmills from being profitable, the Supreme Court ruled that the legal standard for predatory pricing also applies to cases of predatory bidding (*Weyerhaeuser v. Ross-Simmons Hardwood Lumber*, 549 US 312 [2007]). Last year, the Federal Trade Commission (2018a) held hearings to explore monopsony and buyer power in the US economy.

Likewise, Section 7 of the Clayton Act bans mergers that may create or enhance *buyer* market power. For example, in 2018, the Federal Trade Commission challenged the acquisition by Grifols of Biotest, a challenge based in part on the allegation that the acquisition would lessen competition to purchase human plasma, leading to lower fees for people providing plasma (see the case summary at Federal Trade Commission 2018b). Hemphill and Rose (2018) discuss ways to strengthen antitrust enforcement relating to mergers that harm sellers.

The Sherman Act as applied to labor markets prohibits agreements among employers to refrain from competing to hire workers, while the Clayton Act prohibits mergers between employers that may substantially lessen competition in the hiring of workers. Antitrust is generally associated with keeping consumer prices *down* by controlling seller market power, but antitrust applies equally to keeping workers' wages *up* by controlling employer market power.

It seems clear cut that many labor markets depart rather significantly from the textbook model of perfect competition, in which employers are wage takers and face a highly elastic supply of labor. Labor markets are generally defined according to an occupation and a local geographic area (as emphasized by Moretti 2011). With costs of job search and costs of geographical mobility, employers will have some degree of buyer power. Manning (2011) surveys the literature and concludes that "labor markets are pervasively imperfectly competitive." Employers commonly share relationship-specific rents with workers, so employees working at more productive firms earn higher wages (see, for example, Kline et al. 2017; Card et al. 2018).

Some local US labor markets are highly concentrated on the employer side, but that is not the situation for most workers. Azar, Marinescu, and Steinbaum (2017) use data from CareerBuilder.com to calculate labor market concentration in some 8,000 selected labor markets in the United States. They define these labor markets according to occupation and geography, such as "legal secretaries in the Denver area." On average, 20 employers post job vacancies on CareerBuilder.com in a given market in a given quarter. They calculate an employer's share in the labor market based on the number of vacancies listed by that employer at CareerBuilder.com in a given quarter, and they measure market concentration based on the Herfindahl–Hirschman index (HHI) on the employer side of the market. Weighting the geographic markets by population, the overall mean HHI is 1,691, which antitrust economists would classify as moderately concentrated. This method is likely to overestimate labor market concentration, because only about 35 percent of job openings nationally are listed on CareerBuilder.com.

Antitrust enforcement in labor markets has historically been extremely limited. As discussed by Naidu, Posner, and Weyl (2018), this most likely reflects the view that most labor markets are reasonably competitive and that most employers face effective competition to attract and retain workers, combined with a view that some combination of unions, regulations, and lawsuits will help protect workers. That overall conclusion is probably true, but antitrust can still play a role in labor markets in two ways: by considering employer power in labor markets in selected mergers and by addressing anticompetitive agreements in labor markets.

Merger Policy That Considers Labor Markets

A merger that may substantially lessen competition among employers to hire workers is illegal under the Clayton Act. Marinescu and Hovenkamp (2018, 1) note that no merger has ever been blocked on these grounds and infer that "the antitrust law against anticompetitive mergers affecting employment markets is certainly underenforced, very likely by a significant amount." Prager and Schmitt (2019) find that hospital mergers resulting in large increases in concentration in markets for skilled workers, including nurses and pharmacy workers, lead to lower wages.

The Horizontal Merger Guidelines (Department of Justice and Federal Trade Commission 2010, sec. 12) explain how the government evaluates mergers that may enhance buyer power. The government could define a relevant labor market and demonstrate that the merger in question would cause that market to become significantly more concentrated. The merging parties might then try to show that the affected workers have many alternative options for employment. For further details on this type of analysis, see Marinescu and Hovenkamp (2018) and Naidu, Posner, and Weyl (2018).

Two thorny issues are likely to arise if the government begins challenging mergers on the basis of harm to competition in labor markets. First, in cases where the merging parties assert that the merger will reduce their labor costs, the court may need to determine whether to credit these reduced costs as an efficiency gain or instead treat them as the exercise of buyer power in labor markets.[11] Second, if a merger is expected to benefit consumers but harm workers, the court may need to determine whether and how to balance the interests of these two groups. Marinescu and Hovenkamp (2018) argue that under current law, a merger that harms workers by lessening competition in the labor market would not be saved by also offering benefits to consumers.

If the antitrust authorities seriously want to explore the possibility of challenging mergers on the basis of harm to competition in labor markets, developing a quick and efficient means of identifying mergers that involve a significant overlap in plausible labor markets would be a good first step.

Anticompetitive Labor Market Practices: No-Poach and No-Hire Agreements

Section 1 of the Sherman Act prohibits agreements among employers to refrain from competing to hire workers, just as it prohibits traditional cartels among

[11] Anthem's claimed purchasing efficiencies were rejected in *United States v. Anthem*, 855 F.3d 345 (2017).

product-market rivals. This raises questions about no-poach and no-hire agreements that arise in certain labor markets.

In a prominent "no-poach" case, the Department of Justice (2010) sued Adobe, Apple, Google, Intel, Intuit, and Pixar for entering into agreements not to recruit certain workers from each other.[12] When Apple CEO Steve Jobs learned that Google was trying to recruit employees from Apple's Safari team, Jobs threatened Google co-founder Sergey Brin, stating that "if you hire a single one of these people, that means war." In response, Google's CEO Eric Schmidt stopped all efforts at Google to recruit anyone from Apple. When this was conveyed to Apple, Apple reciprocated (Koh 2014). Later, when a Google recruiter contacted an Apple employee, Jobs complained to Schmidt, who apologized and made a public example out of that recruiter, who was terminated within the hour.

The Department of Justice and the Federal Trade Commission later released *Antitrust Guidance for Human Resources Professionals*, stating that "[g]oing forward, the DOJ intends to proceed criminally against naked wage-fixing or no-poaching agreements. These types of agreements eliminate competition in the same irredeemable way as agreements to fix product prices or allocate customers, which have traditionally been criminally investigated and prosecuted as hardcore cartel conduct" (Department of Justice and Federal Trade Commission 2016). Notice that this guidance refers to "*naked* wage-fixing or no-poaching agreements." A no-poach agreement between two or more companies could be justified if those companies are engaged in legitimate joint activity, such as a joint venture to develop new products, and if the no-poach agreement is confined to employees involved in that joint activity, especially if the joint activity involves training these employees or providing them with access to confidential information.

No-hire agreements are common in the franchise sector. Krueger and Ashenfelter (2018) report that in 58 percent of major franchisors' contracts with franchisees, including McDonald's, Burger King, and Jiffy Lube, one franchisee is prohibited from hiring workers from another franchisee in the same chain. They find that no-hire agreements are more common in low-wage, high-turnover industries and have become more common over the past 20 years.

Some limited no-hire provisions of this type could be justified if they provide an incentive for franchisees to invest in workers, giving them human capital that is specific to the franchisor but not to the franchisee. As a result, these agreements are more difficult to challenge under antitrust than are "naked" no-hire agreements. Krueger and Posner (2018) describe a court case involving Jack-in-the-Box in which such a challenge failed. Under the Rule of Reason analysis typically used in antitrust to analyze agreements of this type, two important considerations will be how significantly these agreements restrict the number of employment options available to workers and whether they have depressed wages. A quick look may be sufficient to

[12] I was the chief economist at the Department of Justice when this case was brought.

determine that a no-hire provision has no real efficiency justification and tends to suppress wages.

Conclusion

American antitrust laws date from a time when changes in transportation, communications, and manufacturing technologies generated unprecedented economies of scale, fueling the rise of industrial behemoths. Today, dramatic advances in information technology, combined with globalization, are fueling the growth of large and efficient "superstar firms" that are capturing a growing share of economic activity. The emergence of the tech titans is especially dramatic.

These economic conditions call for a reinvigoration of antitrust enforcement in the United States to promote competition, protect consumers and workers, and spur economic growth. These valuable aims can be achieved by taking a tougher stance toward mergers involving market leaders and by vigilantly preventing dominant firms from engaging in conduct that excludes their rivals. However, moving in that direction is a slow process, requiring the antitrust enforcement agencies to take the lead and convince inertial and possibly skeptical courts to follow. Those who expect dramatic and rapid changes in antitrust will be disappointed, unless new legislation is passed. Likewise, those who expect antitrust to solve problems unrelated to competition will be disappointed. Stronger antitrust enforcement, while needed, is not a substitute for badly needed regulations directed at reducing the political influence of corporations, protecting privacy and data security, and limiting the spread of disinformation.

■ *I thank Jonathan Baker, Aaron Edlin, Joseph Farrell, Steven Salop, Fiona Scott Morton, and Steve Tadelis for their valuable comments on earlier drafts.*

References

Apple. 2019. "The Facts about Parental Control Apps." Apple Press Release, April 28, 2019. https://www.apple.com/newsroom/2019/04/the-facts-about-parental-control-apps/.

Ashenfelter, Orley C., and Daniel S. Hosken. 2010. "The Effect of Mergers on Consumer Prices: Evidence from Five Mergers on the Enforcement Margin." *Journal of Law and Economics* 53(3): 417–66.

Ashenfelter, Orley C., Daniel S. Hosken, and Michael C. Weinberg. 2013. "The Price Effects of a Large Merger of Manufacturers: A Case Study of Maytag-Whirlpool." *American Economic Journal: Economic Policy* 5(1): 239–61.

Ashenfelter, Orley C., Daniel S. Hosken, and Michael C. Weinberg. 2014. "Did Robert Bork Understate the Competitive Impact of Mergers? Evidence from Consummated Mergers." *Journal* *of Law and Economics* 57(S3): S67–100.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2017a. "Concentrating on the Fall of the Labor Share." *American Economic Review* 107(5): 180–85.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2017b. "The Fall of the Labor Share and the Rise of Superstar Firms." NBER Working Paper 23396.

Azar, José, Ioana Marinescu, and Marshall I. Steinbaum. 2017. "Labor Market Concentration." NBER Working Paper 24147.

Azar, José, Martin C. Schmalz, and Isabel Tecu. 2018. "Anticompetitive Effects of Common Ownership." *Journal of Finance* 73(4): 1513–64.

Baker, Jonathan B. 2015. "Taking the Error Out of 'Error Cost' Analysis: What's Wrong with Antitrust's Right." *Antitrust Law Journal* 80(1): 1–38.

Baker, Jonathan B. 2019. *The Antitrust Paradigm: Restoring a Competitive Economy*. Cambridge, MA: Harvard University Press.

Barkai, Simcha. 2017. "Declining Labor and Capital Shares." http://facultyresearch.london.edu/docs/BarkaiDecliningLaborCapital.pdf.

Bessen, James E. 2017. "Information Technology and Industry Concentration." Boston University School of Law, Law and Economics Research Paper 17-41. https://ssrn.com/abstract=3044730.

Blonigen, Bruce A., and Justin R. Pierce. 2016. "Evidence for the Effects of Mergers on Market Power and Efficiency." NBER Working Paper 22750.

Calligaris, Sara, Chiara Criscuolo, and Luca Marcolin. 2018. "Mark-Ups in the Digital Era." OECD Science, Technology and Industry Working Paper 2018/10. https://doi.org/10.1787/4efe2d25-en.

Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. 2018. "Firms and Labor Market Inequality: Evidence and Some Theory." *Journal of Labor Economics* 36(S1): S13–70.

Crémer, Jacques, Yves-Alexandre de Montjoye, and Heike Schweitzer. 2019. *Competition Policy for the Digital Era*. Brussels: European Commission.

Crouzet, Nicolas, and Janice C. Eberly. 2019. "Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles." NBER Working Paper 25869.

Cunningham, Colleen, Florian Ederer, and Song Ma. 2018. "Killer Acquisitions." https://ssrn.com/abstract=3241707.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2018. "The Rise of Market Power and the Macroeconomic Implications." Available at https://sites.google.com/site/deloeckerjan/research.

Decker, Ryan A., John Haltiwanger, Ron S. Jarmin, and Javier Miranda. 2016. "Declining Business Dynamism: What We Know and the Way Forward." *American Economic Review* 106(5): 203–7.

Decker, Ryan A., John Haltiwanger, Ron S. Jarmin, and Javier Miranda. 2017. "Declining Dynamism, Allocative Efficiency, and the Productivity Slowdown." *American Economic Review* 107(5): 322–26.

Demsetz, Harold. 1973. "Industry Structure, Market Rivalry, and Public Policy." *Journal of Law and Economics* 16(1): 1–9.

Department of Justice. 2010. "Justice Department Requires Six High Tech Companies to Stop Entering into Anticompetitive Employee Solicitation Agreements." Department of Justice Press Release 10-1076, September 24, 2010. https://www.justice.gov/opa/pr/justice-department-requires-six-high-tech-companies-stop-entering-anticompetitive-employee.

Department of Justice. 2019. *Criminal Enforcement Trends Charts through Fiscal Year 2018*. https://www.justice.gov/atr/criminal-enforcement-fine-and-jail-charts.

Department of Justice and Federal Trade Commission. 2010. *Horizontal Merger Guidelines*. https://www.justice.gov/atr/file/810276/download.

Department of Justice and Federal Trade Commission. 2016. *Antitrust Guidance for Human Resources Professionals*. https://www.justice.gov/atr/file/903511/download.

Ek, Daniel. 2019. "Consumers and Innovators Win on a Level Playing Field." Spotify Press Release, March 13, 2019. https://newsroom.spotify.com/2019-03-13/consumers-and-innovators-win-on-a-level-playing-field/.

European Commission. 2017. "Antitrust: Commission Fines Google €2.42 Billion for Abusing Dominance as Search Engine by Giving Illegal Advantage to Own Comparison Shopping Service—Factsheet." European Commission Factsheet, June 27, 2017. http://europa.eu/rapid/press-release_MEMO-17-1785_en.htm.

Facebook. 2018. "Response to Six4Three Documents." Facebook Press Release, December 5, 2018. https://newsroom.fb.com/news/2018/12/response-to-six4three-documents/.

Federal Trade Commission. 2018a. "FTC Hearing #2 on Competition and Consumer Protection in the 21st Century: Monopsony and the State of U.S. Antitrust Law." September 21, 2018. https://www.ftc.gov/news-events/events-calendar/2018/09/ftc-hearing-2-competition-consumer-protection-21st-century.

Federal Trade Commission. 2018b. "In the Matter of Grifols, S.A., a Corporation, and Grifols Shared Services North America, Inc., a Corporation." FTC Matter/File No. 181 0081, updated September 18, 2018. https://www.ftc.gov/enforcement/cases-proceedings/181-0081/grifols-sa-grifols-shared-services-north-america-inc-matter.

Federal Trade Commission and Department of Justice. 2018. "Hart-Scott-Rodino Annual Report: Fiscal Year 2017." Available at https://www.ftc.gov/policy/reports/policy-reports/annual-competition-reports.

Federico, Giulio, Fiona Scott Morton, and Carl Shapiro. Forthcoming. "Antitrust and Innovation: Welcoming and Protecting Disruption." Chap. 4 in *Innovation Policy and the Economy*, vol. 20, edited by Josh Lerner and Scott Stern. Chicago: University of Chicago Press. https://www.nber.org/chapters/c14261.

Furman, Jason, Diane Coyle, Amelia Fletcher, Philip Marsden, and Derek McAuley. 2019. *Unlocking Digital Competition: Report of the Digital Competition Expert Panel*. London: Digital Competition Expert Panel.

Furman, Jason, and Peter Orszag. 2018. "A Firm-Level Perspective on the Role of Rents in the Rise of Inequality." In *Toward a Just Society: Joseph Stiglitz and Twenty-First Century Economics*, edited by Martin Guzman, 19–47. New York: Columbia University Press.

Ganapati, Sharat. 2018. "Oligopolies, Prices, Output, and Productivity." https://ssrn.com/abstract=3030966.

Gaynor, Martin. 2018. "Examining the Impact of Health Care Consolidation." Statement before the Committee on Energy and Commerce Oversight and Investigations Subcommittee, US House of Representatives, February 14, 2018. https://docs.house.gov/meetings/IF/IF02/20180214/106855/HHRG-115-IF02-Wstate-GaynorM-20180214.pdf.

Gaynor, Martin, and Robert Town. 2012. "The Impact of Hospital Consolidation: Update." Robert Wood Johnson Foundation, Synthesis Project Policy Brief 9. https://www.rwjf.org/content/dam/farm/reports/issue_briefs/2012/rwjf73261.

Gee, Emily, and Ethan Gurwitz. 2018. *Provider Consolidation Drives Up Health Care Costs: Policy Recommendations to Curb Abuses of Market Power and Protect Patients*. Washington, DC: Center for American Progress. https://cdn.americanprogress.org/content/uploads/2018/12/04110830/Consolidation-HealthCare-Costs.pdf.

Gutiérrez, Germán, and Thomas Philippon. 2018. "How EU Markets Became More Competitive Than US Markets: A Study of Institutional Drift." NBER Working Paper 24700.

Hall, Robert E. 1988. "The Relation between Price and Marginal Cost in U.S. Industry." *Journal of Political Economy* 96(5): 921–47.

Hall, Robert E. 2018. "Using Empirical Marginal Cost to Measure Market Power in the US Economy." NBER Working Paper 25251.

Hemphill, C. Scott, and Nancy L. Rose. 2018. "Mergers That Harm Sellers." *Yale Law Journal* 127(7): 2078–109.

Hemphill, C. Scott, and Philip J. Weiser. 2018. "Beyond *Brooke Group*: Bringing Reality to the Law of Predatory Pricing." *Yale Law Journal* 127(7): 2048–77.

Hovenkamp, Herbert J., and Carl Shapiro. 2018. "Horizontal Mergers, Market Structure, and Burdens of Proof." *Yale Law Journal* 127(7): 1996–2025.

International Monetary Fund. 2019. "The Rise of Corporate Market Power and Its Macroeconomic Effects." Chap. 2 in *World Economic Outlook: Growth Slowdown, Precarious Recovery*. April 2019. Washington, DC: International Monetary Fund.

Kline, Patrick, Neviana Petkova, Heidi Williams, and Owen Zidar. 2017. "Who Profits from Patents? Rent-Sharing at Innovative Firms." Institute for Research on Labor and Employment Working Paper 107-17. http://irle.berkeley.edu/files/2017/Who-Profits-from-Patents.pdf.

Koh, Lucy H. 2014. "In Re: High-Tech Employee Antitrust Litigation, Order Denying Plaintiffs' Motion for Preliminary Approval of Settlements with Adobe, Apple, Google, and Intel." US District Court for the Northern District of California, San Jose Division, Case 11-CV-02509-LHK, Filing 974, August 8, 2014.

Krueger, Alan B., and Orley Ashenfelter. 2018. "Theory and Evidence on Employer Collusion in the Franchise Sector." NBER Working Paper 24831.

Krueger, Alan B., and Eric A. Posner. 2018. "A Proposal for Protecting Low-Income Workers from Monopsony and Collusion." Hamilton Project Policy Proposal 2018-05. http://www.hamiltonproject.org/assets/files/protecting_low_income_workers_from_monopsony_collusion_krueger_posner_pp.pdf.

Kwoka, John. 2014. *Mergers, Merger Control, and Remedies: A Retrospective Analysis of U.S. Policy*. Cambridge, MA: MIT Press.

Levenstein, Margaret C., and Valerie Y. Suslow. 2006. "What Determines Cartel Success?" *Journal of Economic Literature* 44(1): 43–95.

Levenstein, Margaret C., and Valerie Y. Suslow. 2011. "Breaking Up is Hard to Do: Determinants of Cartel Duration." *Journal of Law and Economics* 54(2): 455–92.

Manning, Alan. 2011. "Imperfect Competition in the Labor Market." Chap. 11 in *Handbook of Labor Economics*, vol. 4b, edited by Orley Ashenfelter and David Card, 973–1041. Amsterdam: Elsevier.

Marinescu, Ioana Elena, and Herbert Hovenkamp. 2018. "Anticompetitive Mergers in Labor Markets." University of Pennsylvania Institute for Law and Economics Research Paper 18-8. https://ssrn.com/abstract=3124483.

Moretti, Enrico. 2011. "Local Labor Markets." Chap. 14 in *Handbook of Labor Economics*, vol. 4b, edited by Orley Ashenfelter and David Card, 1237–313. Amsterdam: Elsevier.

Muris, Timothy J., and Jonathan E. Neuchterlein. 2018. "Antitrust in the Internet Era: The Legacy of United States v. A&P." George Mason Law and Economics Research Paper 18-15. https://ssrn.com/abstract=3186569.

Naidu, Suresh, Eric A. Posner, and Glen Weyl. 2018. "Antitrust Remedies for Labor Market Power." *Harvard Law Review* 132(2): 536–601.

O'Sullivan, Donie, and Hadas Gold. 2018. "Facebook Internal Emails Show Zuckerberg Targeting Competitor Vine." *CNN Business*, December 5, 2018. https://www.cnn.com/2018/12/05/media/facebook-six4three-internal-documents-emails/index.html.

Prager, Elena, and Matthew Schmitt. 2019. "Employer Consolidation and Wages: Evidence from Hospitals." Washington Center for Equitable Growth Working Paper. https://ssrn.com/abstract=3391889.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter. 2018. "Diverging Trends in National and Local Concentration." NBER Working Paper 25066.

Salinger, Michael. 1990. "The Concentration-Margins Relationship Reconsidered." *Brookings Papers on Economic Activity* 21(Microeconomics): 297–335.

Schmalensee, Richard. 1989. "Inter-Industry Studies of Structure and Performance." Chap. 16 in *Handbook of Industrial Organization*, vol. 2, edited by Richard Schmalensee and Robert Willig, 951–1009. Amsterdam: Elsevier.

Schmalz, Martin C. 2018. "Common-Ownership Concentration and Corporate Conduct." *Annual Review of Financial Economics* 10(1): 413–48.

Scott Morton, Fiona, Pascal Bouvier, Ariel Ezrachi, Bruno Jullien, Roberta Katz, Gene Kimmelman, A. Douglas Melamed, and Jamie Morgenstern. 2019. *Committee for the Study of Digital Platforms: Market Structure and Antitrust Subcommittee Report. Draft.* Chicago: Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business. https://research.chicagobooth.edu/-/media/research/stigler/pdfs/market-structure—report-as-of-15-may-2019.pdf.

Shapiro, Carl. 2010. "The 2010 Horizontal Merger Guidelines: From Hedgehog to Fox in Forty Years." *Antitrust Law Journal* 77(1): 701–59.

Shapiro, Carl. 2012. "Competition and Innovation: Did Arrow Hit the Bull's Eye?" Chap. 7 in *The Rate and Direction of Inventive Activity Revisited*, edited by Josh Lerner and Scott Stern, 404–10. Chicago: University of Chicago Press.

Shapiro, Carl. 2018. "Antitrust in a Time of Populism." *International Journal of Industrial Organization* 61(1): 714–48.

Syverson, Chad. 2011. "What Determines Productivity?" *Journal of Economic Literature* 49(2): 326–65.

Traina, James. 2018. "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements." https://ssrn.com/abstract=3120849.

Valletti, Tommaso. 2018. "Concentration Trends." Brussels: European Commission. https://www.ecb.europa.eu/pub/conferences/shared/pdf/20180618_ecb_forum_on_central_banking/Valletti_Tommaso_Presentation.pdf.

Valletti, Tommaso M., and Hans Zenger. 2019. "Increasing Market Power and Merger Control." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3387999.

Van Reenen, John. 2018. "Increasing Differences between Firms: Market Power and the Macro-Economy." Centre for Economic Performance Discussion Paper 1576. http://cep.lse.ac.uk/pubs/download/dp1576.pdf.

Werden, Gregory J. 2018. "The 1968 Merger Guidelines: In Praise of Committing to Restraint." *Review of Industrial Organization* 53(3): 445–52.

Wollmann, Thomas G. 2019. "Stealth Consolidation: Evidence from an Amendment to the Hart-Scott-Rodino Act." *American Economic Review: Insights* 1(1): 77–94.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 94–117

The Problem of Bigness: From Standard Oil to Google

Naomi R. Lamoreaux

A number of observers have been sounding the alarm recently about the growth of monopoly power in the US economy. Expressions of concern have issued from all parts of the political spectrum (Langlois 2018), but the most sustained warnings have come from the self-proclaimed "New Brandeisians," a group of scholars for whom the title of Louis Brandeis's famous essay, "A Curse of Bigness" (Brandeis 1914, chap. 8), has become a potent rallying cry. Members of this group claim that Google, Amazon, and other giant tech firms are exploiting blatantly anticompetitive practices to block potential rivals—and getting away with it by manipulating the political system. They are particularly worried that current antitrust orthodoxy, which is preoccupied with the issue of harms to consumers, has left the country all but defenseless against bigness's other ills (Lynn 2010; Khan 2018; Wu 2018; for an overview, see Berk 2018).

The New Brandeisians argue that the country has entered a second Gilded Age, and certainly the concerns they express are much the same as those prompted by the rise of the Standard Oil Trust in that earlier period of turmoil. To late nineteenth-century Americans, Standard was a monster that corrupted politicians and laid waste its competitors. Legislators responded to the mounting pressure to take action by passing antitrust laws at both the state and federal levels beginning in the late 1880s, but these statutes did not prevent other large firms from amassing positions of dominance in most important sectors of the economy over the next two

■ *Naomi R. Lamoreaux is the Stanley B. Resor Professor of Economics and History, Yale University, New Haven, Connecticut, and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Her email address is naomi.lamoreaux@yale.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.94 doi=10.1257/jep.33.3.94

decades. Some of the new giants followed Standard's example and achieved their market power by acquiring competitors. Others grew large by innovating, devising new products or new ways of producing that yielded significant economies of scale. Regardless of the route these firms took to bigness, their sheer size and sudden emergence awoke fears that, unless the government did something fast to prevent it, the giants would entrench themselves by nefarious means.

There was general agreement that Standard had grown large by pursuing anticompetitive practices and should be broken up, and in 1911 the US Supreme Court issued the necessary order. The knottier problem was how to deal with "trusts" (as big businesses were generically called) that acquired their market power by innovating. Although contemporaries tried, following President Theodore Roosevelt's lead, to sort the trusts into "good" and "bad" categories, this exercise in classification turned out to have severe limitations. Because firms always pursue a mix of strategies to "escape from equilibrium," in Levenstein's (2012) apt phrase, deciding which behaviors were pro- and which were anticompetitive was a difficult task. Not only did "good" trusts sometimes resort to "bad" practices to preserve their advantages, but there were many cases in which it was not at all easy to distinguish actions that were anticompetitive in their purpose and effect from those that improved productivity and brought real benefits to consumers. During the so-called Progressive Period—that is, from the turn of the twentieth century to the outbreak of First World War—policymakers struggled with this problem. The solution they arrived at was to write a set of specific prohibitions into the Clayton Antitrust Act of 1914 and simultaneously to create a new regulatory agency, the Federal Trade Commission (FTC), empowered to define and police the boundary.

The boundary between anticompetitive practices and those that enhanced efficiency nonetheless remained difficult to draw. Firms continuously sought new ways to increase their market power, and regulators just as continuously sought new ways to make their efforts illegal. The line between behaviors seen as violating the law and those viewed as legally acceptable shifted back and forth; regulators were excessively vigilant in some periods and excessively lax in others. During the late 1930s, however, in the context of a revival of anti–big business sentiment during the late New Deal, antitrust officials abandoned the attempt to draw the line and instead defined bigness itself as the problem. Their success in inducing the courts to impose antitrust remedies on firms that had not been found guilty of anticompetitive conduct provoked a counterreaction by a group of economic and legal scholars, dubbed the "Chicago School," who in turn prevailed in the courts once economic conditions deteriorated in the 1970s. Advocates of the Chicago School sought to shift the focus of inquiry from whether large firms had market power to whether the market power they possessed had been detrimental to consumers. Like the aggressive trust-busters they opposed, however, they emphatically rejected the preoccupation with conduct that early twentieth-century policymakers had built into the law—just in time for a new wave of giant innovative firms to behave in ways that reanimated those very concerns.

Standard Oil and the Rise of Antitrust

The Standard Oil Company's market share suddenly rose during the 1870s, from about 4 percent of the US petroleum industry to fully 90 percent, sparking the fears that gave birth to the antitrust movement. These fears were not primarily about high prices or harm to consumers. The price of refined petroleum dropped during the 1870s from about 25 cents per gallon to less than 10 cents, much faster than the general price level, and it remained essentially flat in real terms into the twentieth century.[1] Instead, critics focused on Standard's brutal treatment of competitors, particularly its use of secret discounts from railroads to force rivals to sell out. They also worried that Standard's enormous wealth would enable it to wield undue influence over the political system (see especially Tarbell 1904).

These worries had a real basis in fact. As Granitz and Klein (1996) have shown, Standard's rapid rise to dominance owed more to railroad rebates than to any initial advantage in efficiency. Although its refineries were large by the standards of the time, the minimum efficient scale of production was well below Standard's capacity. Nor did Standard benefit from any barriers to entry that might have arisen from superior technology or control of raw-material resources. The industry was competitively structured, with most of the growth in production in the late 1860s and early 1870s coming from new entrants rather than the expansion of existing refineries (Williamson and Daum 1959, chap. 12). Price competition was so intense that producers were driven (unsuccessfully) to collude. After an agreement to form a pool collapsed in 1872, a frustrated John D. Rockefeller, Standard's president, dismissed all such devices as "ropes of sand" (Chandler 1977, p. 321).

Determined to find another way to limit competition in the industry, Rockefeller took advantage of a parallel effort at cartelization that the railroads serving the oil region were undertaking. Like the petroleum refiners, the railroads had been attempting—without success—to restrain price competition, and they hit on a plan that would deploy the refiners as enforcers. The idea was to organize a select group of leading refiners in Cleveland, Pittsburgh, and other production centers into an association called the South Improvement Company, which would then allocate each railroad a share of the business of transporting oil. In return for ensuring that the railroads kept to their allocations, the chosen refiners were granted rebates (discounts) on their own shipments of oil as well as drawbacks (kickbacks) on those of competitors, giving them a significant cost advantage. Although the violent opposition of producers in the oil fields prevented the railroads from actually implementing the plan, there was a period of several months, after the company was formed but before it fell apart, when prospects seemed dire for refineries not

[1] Contemporaries attributed the drop to the expansion of output in the oil fields rather than to any savings from Standard's large-scale operations (New York State 1888, p. 12). Granitz and Klein (1996, p. 30) shared this view, though their data showed that the margin between the prices of crude and refined oil also fell during the 1870s by about 50 percent. Chandler (1977, pp. 321–25) argued that Standard achieved economies of scale in refining and pipeline shipping, though most of the advances he described occurred after the 1870s, when margins were flat.

included in the scheme. Rockefeller took advantage of this period to induce the other Cleveland refiners to sell out to him. As Granitz and Klein (1996) pointed out, companies outside a pool normally have nothing to fear because they can profit from underselling participants. Only the advantages that the South Improvement refineries stood to gain over their competitors can explain why so many rivals ended up selling out to Rockefeller at prices they regarded as below value. Standard emerged from this incident with effective control over the Cleveland segment of the industry and then secretly merged with the participating refiners in other production centers. As a result of these acquisitions and mergers, Standard grew large enough to demand that the railroads continue to grant it rebates, which in turn enabled it to defend its dominance and acquire most of the remaining independent refineries.[2]

That Standard used its resources for political ends is also clear. For example, it is on record as contributing $250,000 to Ohio Republican Party boss Marcus Hanna's fund to defeat William Jennings Bryan, the 1896 Democratic candidate for president (White 2017, p. 846).[3] It also used its influence to try to protect itself from prosecution. A good example was the pressure brought to bear on Ohio's attorney general, David K. Watson, to drop a lawsuit to revoke the company's corporate charter. As was the norm at the time, Ohio law prohibited corporations from holding stock in other companies. Searching for another way to consolidate the company's acquisitions, Standard's lawyers developed a complex type of voting trust, whereby shareholders in the various companies it controlled, including the Standard Oil Company itself, transferred their stock to a board of trustees who voted on their behalf, giving the board powers akin to those of a holding company (Nevins 1953, vol. 1, chap. 21; Hidy and Hidy 1955, pp. 40–49; Williamson and Daum 1959, pp. 466–70). When Watson learned about this arrangement, he filed suit to dissolve the Standard Oil Company on the grounds that participation in the trust violated the terms of its Ohio charter. According to a later attorney general, Watson was repeatedly offered bribes to drop the case. That charge is difficult to substantiate, but there are extant letters from Hanna threatening Watson's political future: "From a party standpoint, interested in the success of the Republican party, and regarding

[2] For a somewhat different explanation of Standard's rise, see Priest (2012). Priest was critical of Granitz and Klein's (1996) account, but his evidence was consistent with their analysis. See also Klein's response (2012). Some historians (for example, Chandler 1977, p. 321) have argued that the rebates Standard received were compensation for the gains in efficiency it offered the railroads. There were surely some such gains, but as Crane (2012) has shown, there are many aspects of the rebate arrangements (especially the drawbacks) that do not fit such a story and can only be explained as anticompetitive. According to a report by the US Bureau of Corporations (1907), Standard continued to receive what were effectively rebates long after they were ostensibly outlawed by the Interstate Commerce Act of 1887. The report spurred Congress to pass new legislation (the Hepburn amendment) that closed the loophole in the law (Johnson 1959, pp. 583–85).

[3] Some of the most nefarious charges were never proven. In one major scandal, for example, Standard stood accused of bribing Ohio legislators to secure a seat in the US Senate for H. B. Payne, the father of the company's treasurer. The charges were compelling enough for the Ohio legislature to conduct an investigation, with troubling though inconclusive results. At that time US senators were chosen by the various state legislatures rather than by the general electorate. See Tarbell (1904, vol. 2, pp. 112–19).

you as in the line of political promotion, I must say that the identification of your office with litigation of this character is a great mistake" (quoted in Bringhurst 1979, p. 14). Watson persevered and won the case in 1892, though rather than revoke the corporation's charter, the Ohio Supreme Court merely required it to withdraw from the trust (*State v. Standard Oil Company*, 49 Ohio St. 137 [1892]). Standard obeyed the letter of the court's order and dissolved the trust, but it preserved its monopoly by moving its corporate domicile to New Jersey and reorganizing as a holding company under that state's newly liberalized general incorporation law (Hidy and Hidy 1955, pp. 219–32; Bringhurst 1979, pp. 12–22).

Standard Oil's success in eliminating competition in the petroleum industry stimulated the formation of similar combinations in a number of other industries, ranging from whisky to lead to sugar to cottonseed oil. As concerns about these new sources of monopoly power rose, most states enacted antitrust laws (more than a dozen before Congress passed the Sherman Act in 1890), and several state attorneys general filed suits to revoke the charters of corporations that participated in trusts (May 1987; Nolette 2012). State initiatives waned, however, when the trusts began to reorganize as New Jersey holding companies, and as a consequence, pressure built on the federal government to step up its own antitrust activities (US Bureau of Corporations 1904; Seager and Gulick 1929; Thorelli 1955). These pressures intensified as a result of the Great Merger Movement of 1896–1904, when about 1,800 firms disappeared into nearly 160 horizontal combinations. Few of the mergers were as dominant in their industries as Standard Oil was in petroleum, but by a conservative estimate, about one-third of them initially had market shares in excess of 70 percent and one-half had more than 40 percent (Lamoreaux 1985, pp. 2–5).

Although the number of Sherman Act prosecutions increased under Presidents Theodore Roosevelt and especially William Howard Taft, federal courts initially found it difficult to apply the law to the so-called tight combinations that took the form of state-chartered corporations. The federal government could act under the Constitution's commerce clause, but it had to tread warily for fear of undermining the states' authority over corporations; once an area of law came within the domain of the commerce clause, state jurisdiction ended (McCurdy 1979; Lamoreaux 1985, pp. 162–69). Eventually, the Supreme Court found a way around that problem in the form of the "Rule of Reason," which it handed down in a pair of landmark decisions breaking up the Standard Oil and the American Tobacco Companies in 1911 (*Standard Oil Company v. United States*, 221 US 1 [1911]; *United States v. American Tobacco*, 221 US 106 [1911]).

According to the Rule of Reason, loose combinations, such as price-fixing agreements among firms, were illegal per se. But tight combinations like Standard could not be held in violation of the Sherman Act by the mere fact of their size. Although "combining . . . so many other corporations, aggregating so vast a capital" gave substance "to the prima facie presumption of intent and purpose" to create a monopoly, the prima facie presumption of intent had to be "made conclusive" by showing that the purpose of the combination was to restrain trade. If that case could be made, the federal government could take action without undermining the states' regulatory powers, for the simple reason that states did not have the authority to charter corporations in violation of federal law. The key then was to demonstrate that the company's domination resulted not from "normal methods of industrial development" but from "new means of combination . . . with the purpose of excluding others from the trade and thus centralizing in the combination a perpetual control" (*Standard Oil v. United States*, p. 75).

This emphasis on "excluding others from the trade" homed in on exactly the behaviors that most worried contemporaries. Although some later commentators have reinterpreted the Rule of Reason as a test of harm to consumers (see especially Bork 1965, 1978), that is a misreading both of the decision and of the context that gave rise to it.[4] In his opinion in the Standard Oil case, Chief Justice Edward Douglass White made no attempt to measure the extent of any damage done to consumers but instead focused on the combines' abusive conduct toward other individuals and firms. "No disinterested mind," he concluded, could survey the evidence about Standard Oil "without being irresistibly driven to the conclusion that the very genius for commercial development and organization which was manifested from the beginning soon begot an intent and purpose . . . to drive others from the field and to exclude them from their right to trade and thus accomplish the mastery which was the end in view." Ticking off the methods Standard used to exclude competitors was enough to demonstrate "a purpose and intent" to monopolize the industry that "we think so certain as practically to cause the subject not to be within the domain of reasonable contention" (*Standard Oil v. United States*, pp. 75–77).

"Good" versus "Bad" Trusts: The Case of Meat-Packing

Once the Supreme Court solved the problem of applying the Sherman Act to state-chartered corporations, the *Standard Oil* case was easy enough to decide; Standard had a virtual monopoly of output in the petroleum industry, and there was abundant evidence that it had acquired its dominant position by predatory means. Other cases, however, posed more difficult issues. What should be done about industries dominated by several large firms (oligopolies) rather than single giant enterprises (monopolies)? What about firms that grew large by innovating—that vanquished competitors because they had developed superior products or because their production processes were more efficient? Although some contemporaries, like Brandeis, regarded bigness itself as a danger, most policymakers thought it was important to distinguish "good" trusts from "bad." Otherwise, regulations designed to prevent anticompetitive behavior might themselves have anticompetitive consequences by constraining innovation (Johnson 1961; Urofsky 1982; McCraw 1984, chap. 3).

[4] Bork's reading of history has been much criticized. For an overview of this literature, see Crane (2014, n3).

The meat-packing industry provides a good example of the difficulties that policymakers faced in distinguishing between good and bad trusts. During the same period that Standard was monopolizing production in the petroleum industry, a small number of very large firms came to dominate meat-packing through a series of innovations that dramatically increased the availability and reduced the price of fresh meat. Many small producers suffered from the resulting gale of creative destruction. Distinguishing their howls of protest from those provoked by unfair competitive practices was not easy, however, in part because of the meat-packers' own behavior. Not content to rely on the advantages generated by their superior efficiency, they resorted to cartels and other types of collusion, triggering a series of antitrust prosecutions and providing critics with abundant evidence of anticompetitive activity.

As late as the 1870s, fresh beef was an expensive and seasonal commodity in Eastern markets. Cattle were shipped live by rail from collection points on the Great Plains and then butchered locally. Not only did shippers have to pay freight charges on substantial parts of the animals that were unsalable, but cattle had to be fed, watered, and otherwise cared for en route and could be transported only when the weather was neither too cold nor too hot. Many entrepreneurs recognized that there would be substantial cost savings from slaughtering cattle in the Midwest and shipping only the dressed beef to Eastern markets, but the first to overcome all the difficulties involved was Gustavus Swift (Chandler 1977, pp. 299–301; Yeager 1981, chap. 3). He collaborated with a refrigeration engineer to design a suitable railroad car and then sank much of his capital into a small fleet. When the railroads, concerned about their substantial investments in cattle cars and feeding stations, refused to carry his cars, he formed an alliance with the Grand Trunk Railroad, the one carrier serving Eastern markets that was not heavily invested in the old technology. Swift bought harvesting rights to ice on the Great Lakes, built a chain of ice stations along the railroad route, and developed partnerships with wholesalers who were willing to distribute his product. Where wholesalers were not cooperative, he competed with them head on—sometimes selling beef directly from his railroad cars at rock-bottom prices. The only firms that could withstand Swift's competition were those with the financial resources to build similar vertically integrated enterprises. By 1887, three had emerged. Together with Swift, they supplied about 85 percent of the interstate market in dressed beef. Other competitors entered over time, but the industry remained highly concentrated, with the top five firms accounting for 75 percent of the interstate market in 1907–1908 and 81 percent in 1916–1917 (Aduddell and Cain 1981, p. 219; Yeager 1981, p. 112).

The meat-packers' innovations enabled consumers to purchase corn-fed beef from the Midwest at prices that undercut the market for the less desirable cattle raised on western ranges. Politically powerful ranchers responded to the decline in their market, as well as the appearance of monopsony buyers for their output, by demanding that Congress investigate and take action against the beef trust. Their voices were joined by those of local butchers and meat wholesalers whose businesses had been hurt by competition from the large packers. As it turned out, there was much to investigate. Although the meat-packers had initially competed vigorously on price, by the late 1880s they were resorting to price-fixing agreements and cartels to keep prices from falling. These efforts continued until 1902, when the Department of Justice secured an injunction against their pool. This avenue of collusion blocked, the three largest producers decided to merge. Although the deal ultimately fell through, in preparation for the consolidation they had each acquired several smaller companies that they then unloaded on a new firm, the National Packing Company, created expressly for that purpose. Jointly owned by the top three meat-packers, National Packing functioned for the next decade as an "evener" that adjusted its level of production as needed to stabilize prices in the industry (Aduddell and Cain 1981, pp. 228–29; Yeager 1981, chap. 6).

President Theodore Roosevelt was one of those who believed in the importance of "discriminating between those combinations which do good and those combinations which do evil" (as quoted in Johnson 1961, p. 418), and he sought discretionary authority to make these kinds of determinations within the executive branch. Congress never granted Roosevelt the powers he sought. In 1903, however, it established the Bureau of Corporations in the new Department of Commerce and Labor.[5] The bureau had no enforcement powers, but it was authorized to conduct investigations of large-scale businesses with the aim of distinguishing good trusts from bad. The idea was that it would use the glare of publicity to discourage bad trusts from pursuing anticompetitive practices.

As worries about the meat-packers' manipulation of the market increased with the formation of the National Packing Company, Congress pressured the Bureau of Corporations to investigate the industry. The bureau complied and issued a report in 1905 that provoked widespread outrage by largely exonerating the companies (US Bureau of Corporations 1905). Adopting an approach remarkably similar to that of the Chicago School today, the bureau focused on the question of whether consumers had been harmed by the meat-packers' actions. The investigators collected data on revenues and costs, from which they concluded that the meat-packers' prices had been reasonable and their profits not excessive. In addition, they bolstered their empirical findings by arguing on theoretical grounds that prices had been held in check by the threat of potential competition, both from new entrants and from local butchers, and that it was unlikely that the packers had engaged in predatory pricing because the structure of the market would have made such a strategy unprofitable. The report did not examine muckrakers' claims that the packers exercised their power in ways that terrorized big and small businesses alike, "[t]o-day . . . compelling a lordly railroad to dismiss its general manager, to-morrow . . . black-listing and ruining some little commission merchant," or that they "thwart[ed] justice and nullif[ied] the laws by the almost undiscoverable methods of partisan politics" (Russell

[5] There was significant congressional opposition to creating the Bureau of Corporations, but President Roosevelt was able to overcome it by strategically releasing a telegram from Standard Oil executives lobbying against the provision (Johnson 1959, p. 577).

1905, pp. 3, 242). It did not address even the most obvious instance of possible collusion—the use the packers made of the National Packing Company—even though the report conceded that the new company "obviously tended to establish a strong community of interest among four of the six leading companies" (US Bureau of Corporations 1905, p. 27). Criticism of the report was so scathing that President Roosevelt, scrambling to get the damage under control, ordered the bureau to publish a supplement (never actually produced) that would provide the public with the answers it demanded (Yeager 1981, pp. 185–90; Murphey 2013, pp. 86–91).

The Bureau of Corporations' report was doubly disastrous because it contaminated the case that federal prosecutors were simultaneously bringing against the meat-packers for violating the injunction against price-fixing. At the request of the federal district attorney in charge of the case, Roosevelt had ordered the bureau to provide the Department of Justice with the data it had collected, but the court ruled that the information could not be used as evidence, effectively sinking the prosecution (Yeager 1981, p. 189; Murphey 2013, p. 91). The Department of Justice continued to seek ways of proceeding against the meat-packers, ultimately filing a criminal indictment in 1910 against the National Packing Company's directors for violating the Sherman Act. That case also failed when, two years later, a jury voted to acquit the men on all the charges (Yeager 1981, chap. 9).[6]

This series of failures nonetheless had a couple of important consequences. First, it prompted the meat-packers to change their behavior. Although they won the criminal case, they learned a crucial lesson from the experience (and from the Supreme Court's articulation the previous year of the Rule of Reason in the *Standard Oil* case): large firms increased their risk of prosecution under the antitrust laws if they interacted with competitors in ways that smacked of cartelization or unfair leverage. The day immediately following the court victory, the three companies that owned National Packing announced that they would dissolve the company and divide up its properties (Yeager 1981, chap. 9). Henceforth they would concentrate on improving their competitive positions by integrating vertically and exploiting economies of scale and scope. By the end of the decade, the five largest firms had acquired controlling interests in the livestock markets handling most of the animals slaughtered in the United States. They had also integrated forward into the wholesale distribution of meat and meat products, as well as by-products of the packing process (Aduddell and Cain 1981).

Second, the failures helped set in motion an effort to revise the Sherman Act. Although the Supreme Court's decision to break up Standard Oil (and American Tobacco) was widely applauded, Chief Justice White's articulation of the Rule of

[6] There were many postmortems of the case in the newspapers. The general consensus was that jurors were reluctant to assess criminal penalties on socially prominent defendants, particularly when only civil charges had been brought in the cases against Standard Oil and American Tobacco, and that they were not able (and indeed did not even try) to follow the technical details of the government's case. See the reports in the *Chicago Daily Tribune* (1912), *Cincinnati Enquirer* (1912), and *New York Tribune* (1912).

Reason sparked worries about how combinations in restraint of trade could ever be considered reasonable (Winerman 2003, pp. 13–15). At the same time, the government's failure to rein in what appeared to be clear instances of collusion, most obviously by the meat-packers, contributed to a general sense that more needed to be done. Although lawmakers in the two major political parties differed on the details, there was broad agreement about the importance of clarifying the meaning of restraints of trade and attempts to monopolize. There was also agreement on the need to create an administrative body that would monitor businesses' adherence to the laws. By mostly lopsided majorities, Congress enacted the Clayton Antitrust and Federal Trade Commission Acts in 1914 (Sklar 1988; Winerman 2003). The first of these laws amended the Sherman Act to prohibit a number of specific practices that had been used for anticompetitive purposes. The second created a new administrative agency tasked with enforcing the antitrust statutes and went further than the Clayton Act by declaring "unfair methods of competition in commerce" to be unlawful (Udell 1957, pp. 14–33).[7]

The meat-packing firms had initially grown large by innovating, but they had responded to the oligopolistic competition that ensued by colluding to control prices and costs. As a result, in the eyes of the public, they had become bad trusts, much like the Standard Oil Company. Indeed, one writer titled his book about them *The Greatest Trust in the World*, claiming that in comparison the Standard Oil Company was "puerile" (Russell 1905). Although after 1912 the packers focused increasingly on improving their competitive position by integrating vertically, they had so damaged their reputations that virtually anything they did was viewed with suspicion by regulators and the media alike. Taking a dim view of their attempts to exploit economies of scale and scope, the new Federal Trade Commission in 1919 charged them with using their dominance of all stages of the production and distribution of meat to monopolize the industry. Even in hindsight, it is difficult to disentangle the efficiency gains that the meat-packers realized through vertical integration from the enhanced ability it gave them both to control prices and exclude competitors. Aduddell and Cain (1981) reviewed the FTC's charges with a skeptical eye and concluded that many could not "be proved or disproved." They conceded, however, that the FTC had uncovered "sufficiently strong evidence to recommend prosecution under both sections I and II of the Sherman Act" (p. 235). In 1920, the packers negotiated a consent decree with the Department of Justice that required them, among other things, to divest themselves of their interests in stockyards and similar facilities (pp. 239–42).

[7] These laws essentially structure antitrust policy to the present day, though there have been some key amendments, most notably, the Robinson–Patman Act of 1936, prohibiting price discrimination; the Celler–Kefauver Act of 1950, allowing the government to block vertical mergers that reduced competition; and the Hart–Scott–Rodino Antitrust Improvements Act of 1976, requiring potentially anticompetitive mergers to be prescreened by the Federal Trade Commission and the Department of Justice.

To Balance or Not to Balance

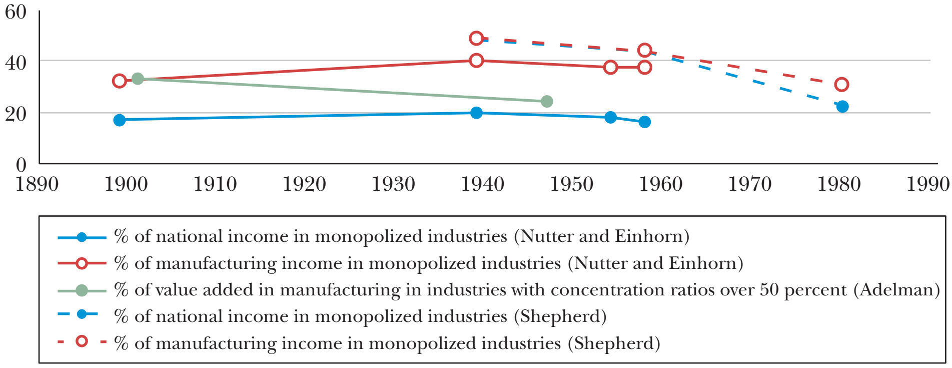
The new antitrust regime put in place in the 1910s meant that firms could no longer acquire monopoly positions in their industries by buying out all their competitors the way Standard Oil had done. The Clayton Act made horizontal mergers illegal when their effect was "substantially to lessen competition or tend to create a monopoly." Firms could still grow large and acquire market power by innovating, so the key policy question became how to prevent businesses that grew large "normally" from turning to anticompetitive practices to preserve their market power, the way the meat-packers had done. The Clayton Act explicitly prohibited certain behaviors, such as tying contracts and discriminatory pricing, and the Federal Trade Commission had broad authority to take action against other conduct regarded as "unfair." As the FTC's case against the meat-packers demonstrated, however, it was often difficult to distinguish actions that increased efficiency from those that forestalled competition. Over time these judgments became even more difficult, as businesses learned how to operate in the new institutional environment without running afoul of the antitrust laws.

The first decade or so of the twentieth century was a difficult period for the giant firms formed during the Great Merger Movement. Although many of these consolidations acquired the bulk of the capacity in their industries, relatively few maintained their dominance for long. Unless they were able to erect barriers to entry (most were not), whenever they tried to raise prices, new firms would enter the market and their market shares would drop (Lamoreaux 1985). Livermore (1935, pp. 90–94) collected earnings data from 1901 to 1932 for 136 mergers that he deemed powerful enough at the time of their formation "to influence market conditions." He found that 37 percent of them were complete failures, while only 44 percent could be regarded as successes. Moreover, those that did not fail had to worry about antitrust prosecution (Bittlingmayer 1993). DuPont was broken up in 1911, shortly after Standard Oil and American Tobacco, and Alcoa signed a consent decree the next year. In the wake of those victories, the Department of Justice launched suits against US Steel and International Harvester, among other companies. Both prosecutions ultimately failed, but they dragged on until the 1920s, absorbing company time and resources.

The firms that survived this shakeout period learned to compete in the new institutional environment by means other than price-cutting. For example, they deployed advertising and other marketing strategies to build brand loyalty, improved their internal operations by integrating backward into raw-material production and forward into distribution, stayed on the technological cutting edge by investing in in-house research and development, and more generally erected barriers to entry in any way they could without inviting antitrust prosecution (Chandler 1977; Lamoreaux 1985). The firms that mastered these lessons dominated their industries for decades. Edwards (1975) has compared the records of the 100 largest firms in the economy in 1903 and 1919. Most of the companies in the 1903 group struggled. Indeed, fully two-thirds were either liquidated within the next two decades or lost

Figure 1

Percentage of National and Manufacturing Income in Monopolized Industries, Selected Years, 1899–1980



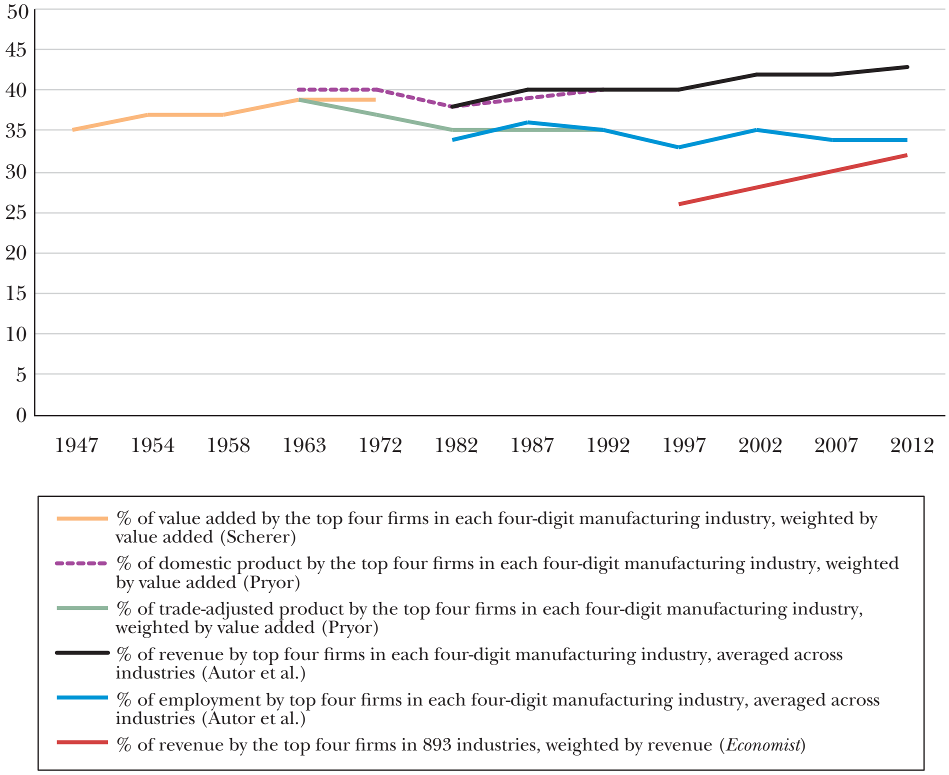
Source: The author, using data from Nutter and Einhorn (1969, pp. 48, 50, 56, 63), Adelman (1951, p. 291), and Shepherd (1982, p. 618). For details of this data and some related data sources, see the Data Appendix available with this article at the *Journal of Economic Perspectives* website.
Note: The authors of the studies cited examined each broad industry or sectoral category to determine whether it was effectively monopolized—that is, dominated by small numbers of large firms.

ground in terms of the real value of their assets. By contrast, most of the firms in the 1919 group were highly successful, with more than 90 percent maintaining at least the real value of their assets a half-century later. Tracking the 100 largest firms in the US economy at various points between 1909 and 1958, Collins and Preston (1961) similarly found that the top firms gradually came to enjoy "an increasing amount of entrenchment of position by virtue of their size" (p. 1001). Over these same decades, moreover, there was remarkably little change in overall levels of economic concentration. Scholars have measured concentration in different ways and over different sets of years, and as a result, their estimates diverge somewhat. But, as can be seen from Figures 1 and 2, there was no clear trend toward increasing (or decreasing) concentration, either in the manufacturing sector or in the economy as a whole.

Intriguingly, even as large firms consolidated their positions, the public's view of them became increasingly accepting. Galambos (1975) analyzed references to big business in a sample of periodicals read by various segments of the middle class over the period 1890–1940 and found that the antipathy of the late nineteenth century had greatly diminished by the interwar period. Auer and Petit (2018) conducted a similar analysis, searching the Proquest database of historical newspapers to find articles that included the word "monopoly." Even though Auer and Petit were selecting on a word with generally negative connotations in American culture, they found that unfavorable mentions dropped from about 75 percent of the total in the late nineteenth century to a little over 50 percent starting in the 1920s.

This process of accommodation was probably furthered by the government's response to popular concerns about the dangers of bigness. In addition to the new

Figure 2
Four-Firm Concentration Ratios, Selected Years, 1947–2012



Source: The author, using data from Scherer (1980, p. 69), Pryor (2001, p. 320), Autor et al. (2017, p. 34), and the *Economist* (2016). For details of this data and some other related data sources, see the Data Appendix available with this article at the *Journal of Economic Perspectives* website.
Notes: All series are for the manufacturing sector, except the one from the *Economist*, which covers the whole economy. Manufacturing's share of total output declined over this period from about 25 to 12 percent.

antitrust laws already discussed, Congress took a first step toward limiting business influence in politics by passing the Tillman Act in 1907, prohibiting corporations from contributing money to political campaigns for national office. The act was a reaction to a particular set of revelations—that large mutual insurance companies were using their members' premiums to lobby for measures that weakened members' protections (Winkler 2004)—but it built on pervasive fears that large-scale businesses were using their vast resources to shape the rules in their favor. By the end of 1908, 19 states had enacted corporate campaign-finance legislation of their own, and they had also begun to restrict lobbying expenditures by corporations (McCormick 1981, p. 266). Congress would write an expanded version of the Tillman law into the Federal Corrupt Practices Act in 1925 (Mager 1976).

The new antitrust regime seems to have been similarly reassuring, even though the 1920s are generally regarded as a period when antitrust enforcement was relatively lax (Cheffins 1989). The Federal Trade Commission got off to an inauspicious start in the early 1920s—most of the complaints it filed were dismissed by the courts—and in the late 1920s it was essentially captured by business interests (Davis 1962). By 1935, however, the agency was showing renewed vitality. The number of complaints it filed increased sharply, its dismissal rate fell to about one-quarter, and it was winning the vast majority of cases that proceeded to judicial review (Posner 1970, p. 382). At the Department of Justice, there was no significant fall-off in the number of cases during the interwar period, with the exception of the early years of the Great Depression. Prosecutors seem to have targeted fewer large firms during the 1920s, but the department's win rate increased from 64 percent in 1920–1924 to 93 percent in 1925–1929 (Posner 1970, pp. 368, 381; Cheffins 1989). Although most antitrust cases still involved horizontal combinations or conspiracies, by the 1930s about one-third of the cases filed by the Department of Justice were targeting abuses of market power, and the FTC's proportion was closer to one-half (Posner 1970, pp. 396, 405, 408).

One of the activities that increasingly concerned antitrust officials during the 1930s was patenting. After World War I, large firms had stepped up both their investments in research and development and their efforts to accumulate patent portfolios. According to surveys conducted by the National Research Council, the number of new industrial research labs grew from about 37 per year between 1909 and 1918 to 74 per year between 1929 and 1936, and research employment in these labs increased by a factor of almost ten between 1921 and 1940 (Mowery and Rosenberg 1989, pp. 62–69). Large firms generated increasing numbers of patents internally, but they also bought them from outside inventors. To measure both streams, Nicholas (2009) collected information on patents assigned at issue during the 1920s to companies that the National Research Council reported as having at least one research lab. Because he was not able to observe assignments that occurred after the patent was granted, his numbers underestimate the stock of patents held by these firms. Nonetheless, he was able to match 17,620 patents to companies listed as having labs in 1927.

The competitive advantages to large firms that broad portfolios of patents could bring, in terms of both what they could achieve technologically and how they could forestall competition, were becoming increasingly apparent—not least to the firms themselves (Reich 1985). As early as the 1920s, valuations on the securities markets began to mirror the size and quality of large firms' patent portfolios (Nicholas 2007). Federal antitrust authorities began to pay attention as well, especially during the late 1930s, when the administration of President Franklin D. Roosevelt displayed renewed interest in the problem of monopoly (Hawley 1966). In 1938, a specially created commission, the Temporary National Economic Committee, launched a three-year investigation into the "Concentration of Economic Power." The Temporary National Economic Committee began its hearings by examining large firms' use of patents to achieve monopoly control, focusing in particular on the automobile and glass industries. In 1939, the committee held a second set of hearings to solicit ideas about how the patent system could be reformed (Hintz 2017). It also commissioned a book-length study by economist Walton Hamilton, *Patents and Free Enterprise* (Hamilton 1941). According to Hamilton, large firms had perverted the patent system. The system's original purpose had been to encourage technological ingenuity, but now large firms were instead deploying patents as barriers to entry and using licensing agreements to divide up the market and limit competition among themselves (Hamilton 1941, pp. 158–63; John 2018).

The Temporary National Economic Committee's patent investigation was headed by Thurman Arnold, assistant attorney general in charge of the Department of Justice's antitrust division. Arnold's views about the abuse of patents were similar to Hamilton's, and at his insistence, the committee's final report recommended compulsory licensing—requiring firms to license their technology at a fair royalty to anyone who wanted to use it. The recommendation went nowhere in Congress (Waller 2004), but Arnold nonetheless pursued it at Justice. As early as 1938, for example, he pushed Alcoa to license a set of its patents as part of an antitrust settlement, and the company agreed in a consent decree entered in 1942. By that time, Arnold had already secured three other compulsory licensing orders, and many more were to follow. Barnett (2018) compiled a complete list of such orders and their terms from 1938 to 1975. By the latter year, the total had risen to 136, one-third of which did not permit the firms to recoup any royalties at all for their intellectual property.

This move against patents was part of a more fundamental shift in antitrust policy that began with Arnold during the late New Deal and then acquired broader intellectual support in the 1950s and 1960s with the spread of what has been called the structure-conduct-performance paradigm, sometimes known as the "Harvard School" (Phillips Sawyer 2019). Most often associated with the work of economist Joe S. Bain (1959), the paradigm's core idea was that the market power of large firms tends to be both self-perpetuating (because size itself confers advantages that operate as barriers to entry) and inimical to consumers (because size is associated with higher profits). The implication was that antitrust authorities should abandon what Bain called their "conduct orientation" and attack the problem of market power directly (p. 607).

Even before the development of this academic rationale, however, a federal appeals court had arrived at essentially the same conclusion in a landmark 1945 decision in the ongoing antitrust suit against Alcoa.[8] Justice Learned Hand's opinion found Alcoa in violation of the Sherman Act because it produced the lion's share (Hand estimated 90 percent) of the country's aluminum ingots and because it was not simply the "passive beneficiary of a monopoly." Alcoa's "crime" was that it continued to behave entrepreneurially and actively seek new business:

[8] The case was heard by the Second Circuit because four of the Supreme Court justices had been associated with prior antitrust litigation against Alcoa and had to recuse themselves. Congress passed a law in 1944 permitting cases where the Supreme Court could not muster a quorum to be certified instead to one of the circuit courts of appeal (Smith 1988).

> True, it stimulated demand and opened new uses for the metal, but not without making sure that it could supply what it had evoked. . . . It was not inevitable that it should always anticipate increases in the demand for ingot and be prepared to supply them. Nothing compelled it to keep doubling and redoubling its capacity before others entered the field (*United States v. Aluminum Company of America*, 148 F.2d 416 [1945], pp. 430–31).

Anticompetitive conduct of the sort that had led to the breakup of Standard Oil was not an issue. As Justice Hand admitted, "We need charge [Alcoa] with no moral derelictions after 1912," the year the company had settled an earlier antitrust suit. "[W]e may assume that all it claims for itself is true"—that the company "won its way by fair means" (pp. 430–31). Alcoa was in violation of the Sherman Act because it was big and successful, not because it had done anything wrong.

The shift in judicial thinking signaled by the Alcoa case stimulated major new antitrust initiatives against AT&T, IBM, and other large innovative firms during the post–World War II era. It also justified the imposition of compulsory licensing orders even in cases where there was no evidence that patents were being used anticompetitively (Barnett 2018). In levying such an order on the United Shoe Machinery Corporation, for example, the court admitted, "Defendant is not being punished for abusive practices respecting patents, for it engaged in none, except possibly two decades ago in connection with the wood heel business. It is being required to reduce the monopoly power it has, not as a result of patents, but as a result of business practices" (*United States v. United Shoe Machinery Corporation*, 110 F. Supp. 295 [1953]). Drawing a line between actions that improved efficiency and those that harmed competition can always be difficult, and it is perhaps especially difficult in the area of patents. But the courts had effectively decided that it was not necessary even to make the attempt.

The new focus on market share in turn provoked a backlash. Economists such as Demsetz (1973, 1974) challenged the structure-conduct-performance paradigm on theoretical grounds, arguing that the observed relationship between size and profits was just a correlation that was more reasonably explained by the likelihood that the most efficient firms would both earn high profits and have a high market share. Legal scholars such as Bork (1965, 1966, 1978) argued that antitrust policy had strayed from its original objective, which was to protect consumers. Like devotees of the structure-conduct-performance paradigm, these "Chicago School" scholars rejected the focus of earlier policymakers on conduct.[9] They simply applied a different test to assess whether a large firm had violated the antitrust laws: instead of measuring the firm's market share, they asked whether the firm had made consumers worse off (Posner 1979).

[9] As Posner (1979) explained, the Chicago School's view of antitrust grew out of a series of studies arguing that predatory pricing, tying contracts, and similar types of bad conduct were economically irrational and so not likely to occur.

The Resurgence of Concerns about Bigness

By the late 1970s, the Chicago School's views were having a major impact on antitrust policy and on the courts (Phillips Sawyer 2019). At this time, inflation was rampant, growth was low, and the US manufacturing sector seemed to be collapsing, transforming what had once been vibrant industrial cities into rust-belt wrecks. The decline had many sources, ranging from external developments such as rising foreign competition to internal problems such as changes in managerial practices that undermined product quality, but regardless it did not seem to be a good time to target the most innovative firms and largest employers in the economy (Hannah 1999; Lamoreaux, Raff, and Temin 2003; Cheffins 2018). Some giant enterprises, including the meat-packers Swift and Armour, disappeared into mergers. Others, like the big three automakers, General Motors, Ford, and Chrysler, struggled to maintain the profitability of their core business; Chrysler survived only with the help of a government bailout. A few successfully reinvented themselves by pursuing different business models. General Electric largely abandoned its consumer electronics business in favor of a strategy of conglomerate mergers. IBM moved away from computer manufacturing and focused instead on business information services and consulting.

Although overall levels of concentration in the economy dipped for a time as a result of these changes (as shown in Figures 1 and 2), they soon began to rise again as new behemoths emerged in the most rapidly growing sectors of the economy, particularly those exploiting cutting-edge computing and information technologies where there were important network externalities (Peltzman 2014; Autor et al. 2017; Gutiérrez and Philippon 2017; Grullon, Larkin, and Michaely forthcoming). In such industries, consumers stood to benefit from using the same products that many others were using, so firms that pulled ahead in the competition quickly acquired dominant market shares. Google, Apple, Amazon, and the other new "superstar" firms (the term comes from Autor et al. 2017) grew primarily by innovating—by offering consumers desirable new products or new ways of buying familiar ones. Nevertheless, their rapid rise set off waves of anxiety about the growth of monopoly power in the American economy reminiscent of the late nineteenth-century reaction to Standard Oil (Lynn 2010; Khan 2018; Wu 2018).

The increase in the market share of these superstar firms does not necessarily mean that the economy had become less competitive, as Carl Shapiro observes in his companion article in this issue. Indeed, evidence to the contrary comes from the simple fact that the identity of the firms singled out as objects of concern changed as technology continued to evolve. In the first decade of the twenty-first century, for example, critics decried Wal-Mart's detrimental effect on competing retailers and the monopsony power it exercised over suppliers and workers (Lichtenstein 2009). By the next decade, however, the spotlight had shifted to Amazon, as internet sales challenged the profitability of brick-and-mortar retailers. According to New Brandeisian Lina Khan (2017, pp. 709–10), for example, Amazon's 46 percent share of e-commerce in the United States does not begin to capture the extent of its dominance. As Khan sees it, Amazon has cut prices and sacrificed profits in a predatory drive to position itself as the indispensable provider of infrastructure services to a broad range of businesses, including those with which it is in competition. Although so far Amazon has generally refrained from exploiting its market power over rivals, it has used its muscle to force down prices charged by its suppliers and by providers of transportation services. In Khan's view, the potential for worse is there, and she argues that the antitrust authorities should take preventive action. However, it is also plausible that ongoing technological progress will give rise to new enterprises that will contest Amazon's hegemony, just as Amazon previously challenged Wal-Mart's.

As the example of the meat-packers makes clear, companies that grow large through innovation are no less likely than those that grow large by merger to turn to anticompetitive practices to maintain their advantages. Microsoft is a recent case in point. The company rose to bigness on the success of its operating system for personal computers and its popular word-processing software. But when faced with new threats to its dominance from computer makers using other operating systems (most notably Apple) and from the growth of the internet, it took steps that even Chicago-influenced antitrust authorities regarded as anticompetitive. According to a lawsuit filed by the Department of Justice, Microsoft promoted the use of its own internet browser by integrating it into its Windows software, negotiating exclusive dealing contracts with internet service providers and software producers, cutting deals with computer makers to install the browser on all the new computers they sold, and threatening those who made similar arrangements with other browser companies with a loss of business. A federal district court found Microsoft in violation of the Sherman Act and ordered the company broken up. An appeals court vacated the breakup order and reversed some of the lower court's findings, but it affirmed other findings and remanded still others for further consideration. Microsoft then settled the case (Cohen 2004). Although the settlement did not completely end the company's legal problems, its executives absorbed the same lessons from the experience that large firms had learned in the early twentieth century: they had to change their ways to avoid antitrust problems.

Once again, the line between actions that improved efficiency and those that aimed "to cut off [rivals'] air supply," as Microsoft's executives were alleged to have threatened (Chandrasekaran 1998), was difficult to draw. Scholars disagreed vehemently about whether Microsoft had transgressed (see, for example, Bresnahan 2001 as well as the symposium in the Spring 2001 issue of this journal, including Klein 2001; Gilbert and Katz 2001; Whinston 2001). Moreover, we can never know what the counterfactual outcome would have been in the absence of litigation. After the settlement, Microsoft's browser sank into obscurity, but so did the competing browsers that were the main beneficiaries of the antitrust action. In 2008, Google introduced Chrome, a new browser that quickly swept away the competition. Ten years later Chrome had a 63 percent share of the global browser market, with Apple's Safari a distant second at 14 percent (Awio Web Services 2018). The browsers involved in the antitrust suit had been completely left in the dust. Would Chrome have been so successful if Microsoft had not been chastened first?

Google now stands accused of using its popular search engine to give preference to its own vertically linked services (Edelman 2015; Phillips Sawyer 2016).

The Federal Trade Commission conducted an investigation of these charges and dismissed them in 2013, deciding that "Google's display of its own content could plausibly be viewed as an improvement in the overall quality of Google's search product" (US Federal Trade Commission 2013, p. 3). By contrast, the European Union's Commissioner for Competition, Margrethe Vestager, found that the biases in Google's search results "artificially divert[ed] traffic from rival comparison shopping services and hinder[ed] their ability to compete" (European Commission 2015). The diversion was detrimental to consumers, the European Commission found, because "users do not necessarily see the most relevant results in response to queries." The European Commission brought formal charges against Google in 2015, and two years later the company was found guilty and fined a record $2.7 billion (as reported in Scott 2017).

Was the Federal Trade Commission too meek, or the European Commission too harsh? The *Wall Street Journal* got a peek behind the curtain of the decision-making process at the FTC when it (accidentally) gained access to scattered pages of an internal FTC staff report on the case through a Freedom of Information Act request (*Wall Street Journal* 2015; Phillips Sawyer 2016, p. 12). Although staff members recommended that the FTC not take action on the charge that Google's search results were biased against competitors, they did encourage the commissioners to sue Google for several other antitrust violations. Moreover, the FTC staff regarded the search engine recommendation to be a "close question": "the evidence paints a complex portrait of a company working toward an overall goal of maintaining its market share by providing the best user experience, while simultaneously engaging in tactics that resulted in harm to many vertical competitors, and likely helped to entrench Google's monopoly power over search and advertising." The recommendation not to move forward with the charge was driven not only by staff members' sense that the line between actions that enhance efficiency and those that are anticompetitive in purpose and effect is difficult to draw, but also by their perception that there was no interest in drawing it in the current antitrust legal environment. Such a determination "would require an extensive balancing of these factors, a task that courts have been unwilling—in similar circumstances—to perform" (*Wall Street Journal* 2015, p. 86 of memorandum).

The Federal Trade Commission promised in its 2013 statement to "remain vigilant and continue to monitor Google for conduct that may harm competition and consumers," and in 2016 it announced that it was expanding its investigation into Google's use of Android to foreclose competition (as reported in Nicas and Kendall 2016).[10] However, critics remain convinced that its focus on consumer welfare is blinding it to the broader range of problems that bigness can entail (Wu 2018).[11] There is renewed concern, moreover, that the tech firms' enormous wealth is giving

[10] As of this writing, it has not yet issued a report, whereas the European Commission recently levied another record fine on Google in the Android case, this time for $5.1 billion (as reported in Satariano and Nicas 2018).

[11] Responding to critics, the Federal Trade Commission announced in 2018 that it would hold public hearings on competition policy, including "whether technology firms are undermining competition." One of the Democratic FTC commissioners also announced that Khan would join his office for a

them undue influence on policy. The Supreme Court's dismantling of restrictions on corporate political contributions (*Citizens United v. Federal Election Commission*, 558 US 310 [2010]) has fueled worries about flows of money that citizens cannot observe. These anxieties have been exacerbated by what they *are* able to observe—the enormous resources that Google and other tech giants have been pouring into lobbying the federal government. Google spent less than $50,000 on lobbying in 2002. In 2017, it spent more than $18 million, and Amazon, Apple, and Facebook were not far behind, with expenditures by these four tech companies totaling nearly $50 million (as reported in Taplin 2017; Bach 2018). In the first three-quarters of 2018, Google spent more on lobbying in Washington ($16.5 million) than any other business corporation, more than the American Medical Association or the American Hospital Association, and considerably more than twice as much as the National Rifle Association (for a list of the firms and organizations that spend the most on lobbying, see Ackley 2018).

The New Brandeisians are raising concerns about the threat that monopoly power poses to the economy and our democracy. These concerns are not new. Indeed, they echo fears aroused by the rise of the Standard Oil Trust and other big businesses at the turn of the last century. Then, as now, the fears were not only about—nor even primarily about—the effect of monopoly on consumers, but rather about the exclusion of competitors from the market and the manipulation of the political system for economic ends. The worries, then as now, had a substantial basis in fact, but they also posed difficult questions of interpretation. How can one determine whether an action was taken to improve efficiency or exclude a rival? What if an action did both?

In response to the rise of Standard Oil and other trusts, lawmakers put in place a complex of institutions that had at their core two basic principles: that firms could grow large by innovating as well as by combining with their competitors, and that even the most innovative enterprises might resort to anticompetitive tactics to preserve their market position. The institutions that early twentieth-century lawmakers created were by no means perfect, but the balance they struck between these competing principles underpinned a long period in which fears of big business abated and large firms learned to stabilize their industries and compete on dimensions other than price without running afoul of the antitrust authorities. Striking the right balance was difficult, however, and policymakers lost their commitment to the effort over the long run, swinging first to the extreme that bigness in itself was bad and needed to be countered, and then to the opposite extreme that bigness was never a problem so long as it brought gains to consumers. Perhaps now would be an opportune time to return to the task of assessing the conduct of large firms. How else can we avoid the twin perils of attacking firms that are large and successful because they are innovative and allowing large, successful firms to block innovative challengers?

few months to advise him on antitrust policy toward Amazon and other tech giants (as reported in McLaughlin 2018).

■ *I'm grateful to Gerben Bakker, Brian Cheffins, Daniel Crane, Richard Daniels, Louis Galambos, Gary Gerstle, Leslie Hannah, Ron Harris, Gordon Hanson, David Lamoreaux, Noam Maggor, Enrico Moretti, Laura Phillips Sawyer, Timothy Taylor, Francesca Trivellato, Heidi Williams, and Mary Yeager for helpful comments and suggestions, and to participants in the Safra Seminar on Law and Ethics at the Tel Aviv University Law School, the Economic History Seminar at the London School of Economics, and the University of Cambridge American History Seminar.*

References

Ackley, Kate. 2018. "Google Still K Street's Top Tech Spender." *Roll Call*, October 24, 2018. https://www.rollcall.com/news/politics/tech-trade-appropriations-keep-k-street-in-business.

Adelman, M. A. 1951. "The Measurement of Industrial Concentration." *Review of Economics and Statistics* 33(4): 269–96.

Aduddell, Robert M., and Louis P. Cain. 1981. "Public Policy toward 'The Greatest Trust in the World.'" *Business History Review* 55(2): 217–42.

Auer, Dirk, and Nicolas Petit. 2018. "Antitrust versus the Press: Two Systems of Belief about Monopoly." https://ssrn.com/abstract=3112150.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen. 2017. "The Fall of the Labor Share and the Rise of Superstar Firms." NBER Working Paper 23396.

Awio Web Services. 2018. "Browser and Platform Market Share: September 2018." W3Counter, September 30, 2018. http://www.w3counter.com/globalstats.php?year=2018&month=9

Bach, Natasha. 2018. "Google Outspent Every Other Company on Washington Lobbying Last Year." *Fortune*, January 24, 2018. http://fortune.com/2018/01/24/google-facebook-amazon-apple-lobbying-efforts/.

Bain, Joe S. 1959. *Industrial Organization.* New York: John Wiley and Sons.

Barnett, Jonathan M. 2018. "The Great Patent Grab." Unpublished.

Berk, Gerald. 2018. "The New Brandeisians: Rethinking Antitrust in the Open Markets Institute." Unpublished.

Bittlingmayer, George. 1993. "The Stock Market and Early Trust Enforcement." *Journal of Law and Economics* 36(1): 1–32.

Bork, Robert H. 1965. "The Rule of Reason and the Per Se Concept: Price Fixing and Market Division." *Yale Law Journal* 74(5): 775–847.

Bork, Robert H. 1966. "Legislative Intent and the Policy of the Sherman Act." *Journal of Law and Economics* 9(1): 7–48.

Bork, Robert H. 1978. *The Antitrust Paradox: A Policy at War with Itself.* New York: Basic Books.

Brandeis, Louis D. 1914. *Other People's Money and How the Bankers Use It.* New York: Frederick A. Stokes.

Bresnahan, Timothy F. 2001. "A Remedy That Falls Short of Restoring Competition." *Antitrust* 16(Fall): 67–71.

Bringhurst, Bruce. 1979. *Antitrust and the Oil Monopoly: The Standard Oil Cases, 1890–1911.* Westport, CT: Greenwood.

Chandler, Alfred D. Jr. 1977. *The Visible Hand: The Managerial Revolution in American Business.* Cambridge, MA: Harvard University Press.

Chandrasekaran, Rajiv. 1998. "Microsoft Attacks the Credibility of Intel Exec." *Washington Post,* November 13, 1998, B1.

Cheffins, Brian. 1989. "The Development of Competition Policy, 1890–1940: A Re-evaluation of a Canadian and American Tradition." *Osgoode Hall Law Journal* 27(3): 449–90.

Cheffins, Brian R. 2018. *The Public Company Transformed.* New York: Oxford University Press.

Chicago Daily Tribune. 1912. "Packers Free." March 27, 1912, 1.

Cincinnati Enquirer. 1912. "Jurors Muddled by Figures." March 27, 1912, 5.

Cohen, Amanda. 2004. "Surveying the Microsoft Antitrust Universe." *Berkeley Technology Law Journal* 19(1): 333–64.

Collins, Norman R., and Lee E. Preston. 1961. "The Size Structure of the Largest Industrial Firms, 1909–1958." *American Economic Review* 51(5): 986–1011.

Crane, Daniel A. 2012. “Were Standard Oil’s Rebates and Drawbacks Cost Justified?” *Southern California Law Review* 85(3): 559–72.

Crane, Daniel A. 2014. “The Tempting of Antitrust: Robert Bork and the Goals of Antitrust Policy.” *Antitrust Law Journal* 79(3): 835–53.

Davis, G. Cullom. 1962. “The Transformation of the Federal Trade Commission, 1914–1929.” *Mississippi Valley Historical Review* 49(3): 437–55.

Demsetz, Harold. 1973. “Industry Structure, Market Rivalry, and Public Policy.” *Journal of Law and Economics* 16(1): 1–9.

Demsetz, Harold. 1974. “Two Systems of Belief about Monopoly.” In *Industrial Concentration: The New Learning*, edited by Harvey J. Goldschmid, H. Michael Mann, and J. Fred Weston, 164–83. Boston: Little, Brown.

Economist. 2016. “Corporate Concentration: The Creep of Consolidation across America’s Corporate Landscape.” March 24, 2016. https://www.economist.com/graphic-detail/2016/03/24/corporate-concentration.

Edelman, Benjamin. 2015. “Does Google Leverage Market Power through Tying and Bundling?” *Journal of Competition Law and Economics* 11(2): 354–400.

Edwards, Richard C. 1975. “Stages in Corporate Stability and the Risks of Corporate Failure.” *Journal of Economic History* 35(2): 428–57.

European Commission. 2015. “Antitrust: Commission Sends Statement of Objections to Google on Comparison Shopping Service; Opens Separate Formal Investigation on Android.” Press Release IP/15/4780. http://europa.eu/rapid/press-release_IP-15-4780_en.htm.

Galambos, Louis. 1975. *The Public Image of Big Business in America, 1880–1940: A Quantitative Study in Social Change.* Baltimore: Johns Hopkins University Press.

Gilbert, Richard J., and Michael L. Katz. 2001. “An Economist’s Guide to *U.S. v. Microsoft*.” *Journal of Economic Perspectives* 15(2): 25–44.

Granitz, Elizabeth, and Benjamin Klein. 1996. “Monopolization by ‘Raising Rivals’ Costs’: The Standard Oil Case.” *Journal of Law and Economics* 39(1): 1–47.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely. Forthcoming. “Are U.S. Industries Becoming More Concentrated?” *Review of Finance.* https://ssrn.com/abstract=2612047.

Gutiérrez, Germán, and Thomas Philippon. 2017. “Declining Competition and Investment in the U.S.” NBER Working Paper 23583.

Hamilton, Walton. 1941. *Patents and Free Enterprise.* Temporary National Economic Committee Monograph 31. Washington, DC: Government Printing Office.

Hannah, Leslie. 1999. “Marshall’s ‘Trees’ and the Global ‘Forest’: Were ‘Giant Redwoods’ Different?” In *Learning by Doing in Markets, Firms, and Countries*, edited by Naomi R. Lamoreaux, Daniel M. G. Raff, and Peter Temin, 253–86. Chicago: University of Chicago Press.

Hawley, Ellis W. 1966. *The New Deal and the Problem of Monopoly: A Study in Economic Ambivalence.* Princeton, NJ: Princeton University Press.

Hidy, Ralph W., and Muriel E. Hidy. 1955. *Pioneering in Big Business, 1882–1911: History of Standard Oil Company (New Jersey).* New York: Harper and Brothers

Hintz, Eric S. 2017. “The ‘Monopoly’ Hearings, Their Critics, and the Limits of Patent Reform in the New Deal.” In *Capital Gains: Business and Politics in Twentieth-Century America*, edited by Richard R. John and Kim Phillips-Fein, 61–79. Philadelphia: University of Pennsylvania Press.

John, Richard R. 2018. “Patents and ‘Free Enterprise’: The TNEC Reconsidered.” Unpublished.

Johnson, Arthur M. 1959. “Theodore Roosevelt and the Bureau of Corporations.” *Mississippi Valley Historical Review* 45(4): 571–90.

Johnson, Arthur M. 1961. “Antitrust Policy in Transition, 1908: Ideal and Reality.” *Mississippi Valley Historical Review* 48(3): 415–34.

Khan, Lina M. 2017. “Amazon’s Antitrust Paradox.” *Yale Law Journal* 126(3): 710–805.

Khan, Lina. 2018. “The New Brandeis Movement: America’s Antimonopoly Debate.” *Journal of European Competition Law and Practice* 9(3): 131–32.

Klein, Benjamin. 2001. “The Microsoft Case: What Can a Dominant Firm Do to Defend Its Market Position?” *Journal of Economic Perspectives* 15(2): 45–62.

Klein, Benjamin. 2012. “The ‘Hub-and-Spoke’ Conspiracy That Created the Standard Oil Monopoly.” *Southern California Law Review* 85(3): 459–98.

Lamoreaux, Naomi R. 1985. *The Great Merger Movement in American Business, 1895–1904.* Cambridge: Cambridge University Press.

Lamoreaux, Naomi R., Daniel M. G. Raff, and Peter Temin. 2003. “Beyond Markets and Hierarchies: Toward a New Synthesis of American Business History.” *American Historical Review* 108(2): 404–33.

Langlois, Richard N. 2018. “Hunting the Big Five: Twenty-First Century Antitrust in Historical Perspective.” https://ssrn.com/abstract=3124356.

Levenstein, Margaret C. 2012. “Escape from Equilibrium: Thinking Historically about Firm Responses to Competition.” *Enterprise and Society* 13(4): 710–28.

Lichtenstein, Nelson. 2009. *The Retail Revolution: How Wal-Mart Created a Brave New World of*

Business. New York: Metropolitan Books.

Livermore, Shaw. 1935. "The Success of Industrial Mergers." *Quarterly Journal of Economics* 50(1): 68–96.

Lynn, Barry C. 2010. *Cornered: The New Monopoly Capitalism and the Economics of Destruction.* Hoboken, NJ: John Wiley and Sons.

Mager, T. Richard. 1976. "Past and Present Attempts by Congress and the Courts to Regulate Corporate and Union Campaign Contributions and Expenditures in the Election of Federal Officials." *Southern Illinois University Law Journal* 1(2): 338–99.

May, James. 1987. "Antitrust Practice and Procedure in the Formative Era: The Constitutional and Conceptual Reach of State Antitrust Law, 1880–1918." *University of Pennsylvania Law Review* 135(3): 495–593.

McCormick, Richard L. 1981. "The Discovery that Business Corrupts Politics: A Reappraisal of the Origins of Progressivism." *American Historical Review* 86(2): 247–74.

McCraw, Thomas K. 1984. *Prophets of Regulation: Charles Francis Adams, Louis D. Brandeis, James M. Landis, Alfred E. Kahn.* Cambridge, MA: Harvard University Press.

McCurdy, Charles W. 1979. "The *Knight* Sugar Decision of 1895 and the Modernization of American Corporation Law, 1869–1903." *Business History Review* 53(3): 304–42.

McLaughlin, David. 2018. "Amazon Antitrust Critic Joins FTC as Agency Sets Sights on Tech." *Bloomberg,* July 9, 2018. https://www.bloomberg.com/news/articles/2018-07-09/amazon-antitrust-critic-joins-ftc-as-agency-sets-sights-on-tech.

Mowery, David C., and Nathan Rosenberg. 1989. *Technology and the Pursuit of Economic Growth.* Cambridge: Cambridge University Press.

Murphey, William. 2013. "Theodore Roosevelt and the Bureau of Corporation: Executive-Corporate Cooperation and the Advancement of the Regulatory State." *American Nineteenth Century History* 14(1): 73–111.

Nevins, Allan. 1953. *Study in Power: John D. Rockefeller, Industrialist and Philanthropist.* 2 vols. New York: Scribner.

New York State. 1888. *Report of the Committee on General Laws on the Investigation Relative to Trusts.* Albany, NY: Troy Press.

New York Tribune. 1912. "Jury Acquits Packers." March 27, 1912, 5.

Nicas, Jack, and Brent Kendall. 2016. "FTC Extends Probe into Google's Android." *Wall Street Journal,* April 26, 2016. https://www.wsj.com/articles/ftc-extends-probe-into-googles-android-1461699217.

Nicholas, Tom. 2007. "Stock Market Swings and the Value of Innovation, 1908–1929." In *Financing Innovation in the United States, 1870 to the Present,* edited by Naomi R. Lamoreaux and Kenneth L. Sokoloff, 217–45. Cambridge, MA: MIT Press.

Nicholas, Tom. 2009. "Spatial Diversity in Invention: Evidence from the Early R&D Labs." *Journal of Economic Geography* 9(1): 1–31.

Nolette, Paul. 2012. "Litigating the 'Public Interest' in the Gilded Age: Common Law Business Regulation by Nineteenth-Century State Attorneys General." *Polity* 44(3): 373–99.

Nutter, G. Warren, and Henry Adler Einhorn. 1969. *Enterprise Monopoly in the United States, 1899–1958.* New York: Columbia University Press.

Peltzman, Sam. 2014. "Industrial Concentration under the Rule of Reason." *Journal of Law and Economics* 57(S3): S101–20.

Phillips Sawyer, Laura. 2016. "Google in Europe: Competition Policy in the Digital Era." Harvard Business School Case 9-717-004. https://hbsp.harvard.edu/product/717004-PDF-ENG.

Phillips Sawyer, Laura. 2019. "U.S. Antitrust Law and Policy in Historical Perspective." Harvard Business School Working Paper Series 19-110. https://www.hbs.edu/faculty/Pages/item.aspx?num=56116.

Posner, Richard A. 1970. "A Statistical Study of Antitrust Enforcement." *Journal of Law and Economics* 13(2): 365–419.

Posner, Richard A. 1979. "The Chicago School of Antitrust Analysis." *University of Pennsylvania Law Review* 127(4): 925–48.

Priest, George L. 2012. "Rethinking the Economic Basis of the Standard Oil Refining Monopoly: Dominance against Competing Cartels." *Southern California Law Review* 85(3): 499–557.

Pryor, Frederic L. 2001. "New Trends in U.S. Industrial Concentration." *Review of Industrial Organization* 18(3): 301–26.

Reich, Leonard S. 1985. *The Making of American Industrial Research: Science and Business at GE and Bell, 1876–1926.* New York: Cambridge University Press.

Russell, Charles Edward. 1905. *The Greatest Trust in the World.* New York: Ridgway-Thayer.

Satariano, Adam, and Jack Nicas. 2018. "E.U. Fines Google $5.1 Billion in Android Antitrust Case." *New York Times,* July 18, 2018. https://www.nytimes.com/2018/07/18/technology/google-eu-android-fine.html.

Scherer, Frederic M. 1980. *Industrial Market Structure and Economic Performance.* 2nd ed. Chicago: Rand McNally.

Scott, Mark. 2017. "Google Fined Record $2.7 Billion in E.U. Antitrust Ruling." *New York Times,* June 27, 2017. https://www.nytimes.

com/2017/06/27/technology/eu-google-fine.html.

Seager, Henry R., and Charles A. Gulick Jr. 1929. *Trust and Corporation Problems*. New York: Harper and Brothers.

Shepherd, William G. 1982. "Causes of Increased Competition in the U.S. Economy, 1939–1980." *Review of Economics and Statistics* 64(4): 613–26.

Sklar, Martin J. 1988. *The Corporate Reconstruction of American Capitalism, 1890–1916: The Market, the Law, and Politics*. Cambridge: Cambridge University Press.

Smith, George David. 1988. *From Monopoly to Competition: The Transformation of Alcoa, 1888–1986*. Cambridge: Cambridge University Press.

Taplin, Jonathan. 2017. "Why Is Google Spending Record Sums on Lobbying Washington?" *Guardian*, July 30, 2017. https://www.theguardian.com/technology/2017/jul/30/google-silicon-valley-corporate-lobbying-washington-dc-politics.

Tarbell, Ida M. 1904. *The History of the Standard Oil Company*. 2 vols. New York: McClure, Phillips and Co.

Thorelli, Hans B. 1955. *The Federal Antitrust Policy: Origination of an American Tradition*. Baltimore: Johns Hopkins Press.

Udell, Gilman G., comp. 1957. *Antitrust Laws with Amendments, 1890–1956*. Washington, DC: Government Printing Office.

Urofsky, Melvin I. 1982. "Proposed Federal Incorporation in the Progressive Era." *American Journal of Legal History* 26(2): 160–83.

US Bureau of Corporations. 1904. *Report of the Commissioner of Corporations*. H. Doc. 165, 58th Cong., 3d Sess. Washington, DC: Government Printing Office.

US Bureau of Corporations. 1905. *Report of the Commissioner of Corporations on the Beef Industry*. H. Doc. 382, 58th Cong., 3d Sess. Washington, DC: Government Printing Office.

US Bureau of Corporations. 1907. *Report of the Commissioner of Corporations on the Petroleum Industry*, parts I and II. Washington, DC: Government Printing Office.

US Federal Trade Commission. 2013. "Statement of the Federal Trade Commission Regarding Google's Search Practices—in the Matter of Google Inc., FTC File Number 111-0163." https://www.ftc.gov/public-statements/2013/01/statement-federal-trade-commission-regarding-googles-search-practices.

Wall Street Journal. 2015. "The FTC Report on Google's Business Practices: Scan the Document from the U.S. Antitrust Investigation of the Internet Giant." March 24, 2015. http://graphics.wsj.com/google-ftc-report/.

Waller, Spencer Weber. 2004. "The Antitrust Legacy of Thurman Arnold." *St. John's Law Review* 78(3): 569–613.

Whinston, Michael D. 2001. "Exclusivity and Tying in *U.S. v. Microsoft*: What We Know, and Don't Know." *Journal of Economic Perspectives* 15(2): 63–80.

White, Richard. 2017. *The Republic for Which It Stands: The United States during Reconstruction and the Gilded Age, 1865–1896*. New York: Oxford University Press.

Williamson, Harold F., and Arnold R. Daum. 1959. *The American Petroleum Industry: The Age of Illumination, 1859–1899*. Evanston, IL: Northwestern University Press.

Winerman, Marc. 2003. "The Origins of the FTC: Concentration, Cooperation, Control, and Competition." *Antitrust Law Journal* 71(1): 1–97.

Winkler, Adam. 2004. "'Other People's Money': Corporations, Agency Costs, and Campaign Finance Law." *Georgetown Law Journal* 92(5): 871–938.

Wu, Tim. 2018. *The Curse of Bigness: Antitrust in the New Gilded Age*. New York: Columbia Global Reports.

Yeager, Mary. 1981. *Competition and Regulation: The Development of Oligopoly in the Meat Packing Industry*. Greenwich, CT: JAI Press.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 118–143

How Market Design Emerged from Game Theory: A Mutual Interview

Alvin E. Roth and Robert B. Wilson

We go back a fairly long way. Al was a PhD student at Stanford in Operations Research from 1971 to '74, and Bob was his dissertation advisor. Game theory was also young in those days; its offspring, mechanism design, was even younger; and practical market design by economists was not yet on the horizon.

To jog our memories about the history and development of game theory and how it shaped and was reshaped by market design, we interviewed each other over coffee during Fall 2018.[1] We also touched on what we think has been learned about markets and marketplaces by trying to design them.

What emerged from our discussion is that, when we learned game theory, games were modeled either in terms of the strategies available to the players ("noncooperative game theory") or in terms of the outcomes that could be attained by coalitions of players ("cooperative game theory"), and these were viewed as models appropriate for different kinds of games. In either case, the particular model was viewed as a mathematical object that could be viewed in its entirety by the theorist. Market design, however, has come to view these models as complementary approaches for

[1] We later added publication dates for the work to which we refer, and each of us inserted footnotes to our own comments where additional background seemed useful.

■ *Alvin E. Roth is the Craig and Susan McCaw Professor of Economics and Robert B. Wilson is the Adams Distinguished Professor of Management, Emeritus, Graduate School of Business, both at Stanford University, Stanford, California. Their email addresses are alroth@stanford.edu and rwilson@stanford.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.118 doi=10.1257/jep.33.3.118

examining different ways in which marketplaces operate within their economic environment. And, because that environment can be complex, there will be aspects of the game that are not entirely observable.

Mathematical models themselves play a less heroic, stand-alone role in market design than in the theoretical mechanism design literature. A lot of other kinds of investigation, communication, and persuasion play a role in crafting a workable design and in helping it to be adopted and implemented, and then maintained and adapted.

How Did Game Theory Look When You Began to Learn It, and Teach It?

Wilson: Before 1960, basic concepts of strategic analysis were established but had slight influence on economics. Studies of parlor games (for example, Borel 1921, 1953) influenced von Neumann's early work on minmax solutions of constant-sum two-player games. Von Neumann and Morgenstern (1944) showed existence of such solutions for all such games and then proposed a solution of cooperative games.[2] Nash's (1950a, 1951) definition of equilibrium for noncooperative games offered an alternative approach. The main applications to noncooperative contexts were military and about zero-sum two-player games until Schelling's (1960) broad view of "the strategy of conflict," which was nontechnical but informed by game theory and widely read. Axiomatic cooperative theory advanced via the "value" introduced by Shapley (1953) and the "bargaining solution" by Nash (1950b), and these were often invoked in theoretical economic models.[3]

Most influential for economists was Luce and Raiffa's (1957) book-length critique of game theory's potential for advancements in the social sciences; it was guardedly optimistic, with severe criticisms, widely read, and influenced a generation of scholars. Hayek's (1945) article on "the use of knowledge in society" set the stage for much-later use of models from game theory. He interpreted markets as mechanisms for eliciting preferences and equating marginal rates of substitution among diverse agents with local information about production and consumption opportunities. In the early 1960s, I read most of Luce and Raiffa (1957) and portions of von Neumann and Morgenstern (1944) and Karlin (1959).

Roth: I also found Luce and Raiffa much easier to read than von Neumann and Morgenstern.

[2] Their book was heralded as the future of economics in reviews by Hurwicz (1945), Marschak (1946), Copeland (1945), and Wald (1947). Marschak concluded: "Ten more such books and the progress of economics is assured." Copeland wrote: "Posterity may regard this book as one of the major scientific achievements of the first half of the twentieth century." It took decades for these prospects to be realized, albeit in a form rather different than von Neumann and Morgenstern envisaged initially.

[3] An excellent complement to our discussion here is Myerson's (1999) history of Nash's contributions and their subsequent impact in economic theory.

Wilson: As an MBA student in 1960, I wrote a class report on how to bid in an auction that got a failing grade because it was not "managerial." My studies with Howard Raiffa were focused on decision theory, but discussions with Jacob Marschak and Lloyd Shapley turned me to game theory, initially to generalize the Lemke and Howson (1964) algorithm to find Nash equilibria of N-player games, then in papers on auctions, then in advising the brilliant Armando Ortega-Reichert (1969) on his dissertation about auctions that went far beyond Vickrey (1961). Social choice theory and auctions were main interests until 1978, although in 1968 I developed an MBA course on "competitive strategies" with broad coverage, and a PhD course on "multiperson decision theory."

Roth: When I began grad school at Stanford in 1971, there was no course offered in game theory. But Michael Maschler visited in academic year '72–73 and offered one.[4]

In those days, game theory was thought of as being divided into two parts: cooperative and noncooperative. These were entirely separate theories, differently formulated and thought to apply to different economic environments—namely, those with and without binding agreements. The idea was that for cooperative games, we would study what (binding) agreements rational agents would reach. For noncooperative games, we would study Nash (1950a) equilibria, interpreted either as the result of players' independent optimization in the light of others' presumed rationality, or as the agreements they could reach (in the absence of ways to enforce agreements) that would be self-enforcing in the sense that no player had an incentive to break the agreement if others were expected to follow it when they chose their strategies.

This division of game theory into two parts had its origins in von Neumann and Morgenstern (1944), although some of the particular ideas, interpretations, and models (including Nash's formulation of equilibria) came later.[5]

Also inherited from von Neumann and Morgenstern was that the goal of game theory should be to find the "solution" to each class of games, that would "solve" each kind of theory. They attached great importance to the idea that a solution, when found, would apply to all games in (at least) a very broad class, and therefore that an important property of prospective solutions should be that they should exist for all games. Indeed, an existence proof was often regarded as the main contribution of Nash (1950a), rather than his novel formulation of strategic equilibrium.

There were two complementary models of noncooperative games: the "normal" or strategic form of the game represented as an n-dimensional matrix when n players were involved, and Kuhn's (1950, 1953) formulation of extensive-form games. The extensive form is a tree with branches representing the actions

[4] Jerusalem was the center of the game theory world at that time, with a thriving group that had grown up around Bob Aumann, which included Maschler and a growing group of top students at the Hebrew University and later also at Tel Aviv, which by 1972 already included Bezalel Peleg, David Schmeidler, and Shmuel Zamir.

[5] There weren't any up-to-date advanced textbooks in the 1970s, although Owen (1968) provided an introduction, and so Maschler taught from papers and his own notes.

available to particular players as a consequence of actions taken earlier in the tree, and "information sets" of nodes which indicated what a player knew about earlier decisions when it was his turn to move (with all the nodes in a given information set being indistinguishable to that player at the time when the choice of action was demanded). In this formulation, a strategy for a player was a complete plan of action: a function that specified for each of that player's information sets, what action he would take if the game reached that information set. The more compact "normal" or "strategic" form of the game came to be understood as specifying the (expected utility) payoffs that each player would receive as a consequence of each possible combination of strategy choices by the players.

The "solution concept" for predicting players' choices of strategies was Nash equilibrium.[6] Selten (1965) had already introduced, in German, the subgame perfect refinement of Nash equilibrium in the extensive form, but most English-speaking game theorists learned about that only from his 1975 article that introduced what came to be called "trembling hand perfection" as a further refinement. For many years, the search for increasingly powerful refinements was a hot topic in game theory. The idea behind refinements was very closely related to the conception of each game as being perfectly captured by its extensive or strategic form: if we knew everything about a game, then perhaps we could deduce from first principles which of the multiplicity of equilibria would be the one that would be picked by perfectly rational agents who knew one another to be perfectly rational.[7] While this was never achieved, some refinements, including various notions of perfection, and of sequential rationality (Kreps and Wilson 1982), have become useful tools for modern game theorists, and new refinements continue to be proposed and explored (for example, Milgrom and Mollner 2018; Myerson and Weibull 2015), because many if not most games have a plethora of Nash equilibria.

Harsanyi (1967–68) extended the extensive-form model with common knowledge to games of incomplete information, in which there was an initial common knowledge move by Nature that produced "types" of players who each knew his own type but knew only the distribution of other players' types. The idea that everything about the structure of the game (including the rationality of all the players) was known to all the players was made clearer by specifying precisely what it meant for something to be common knowledge (made formal by Aumann 1976 independently of, but in the spirit of, Lewis 1969).

[6] The awkward term "solution concept" was widely used once it became clear that there were not readily going to be any perfect "solutions" forthcoming, although the word "perfect" would emerge as a term in the equilibrium refinement literature, which continued to seek a definitive solution for noncooperative games.

[7] The hope was that the "right" refinement would be a subset of all the others, possibly a unique equilibrium for each game that captured most fully the perfect rationality of all the players. However, refinements of equilibria turned out to be more like onions than like olives: applying all of the attractive refinement principles to peel away imperfect equilibria did not yield an irreducible center, but rather nothing at all. *Wilson:* I view the axioms in Govindan and Wilson (2012) that characterize Mertens (1989) stable sets as a surviving core.

I'll come back to this, as the idea that the entire game—all the strategies available to all the participants—was common knowledge among the participants and completely known to the theorist seeking to analyze the game, was one of the features of early game theory that had to be overcome for practical market design to develop. Indeed, the idea that practical design should not depend on unrealistic common knowledge assumptions (or on assumed knowledge involving too much detail of players' private information) is widely known as the "Wilson doctrine," after Wilson (1987).

Wilson: At issue was whether the fine detail of game-theoretic models provided new economic insights.[8] Is it sufficient to assume that markets clear at whatever prices are required to do it, or should one examine institutional arrangements for eliciting demands and establishing prices, or even the informational content of prices as suggested by Hayek? It was hard to give up the beautiful welfare theorems implied by "perfect competition" if one were to take account of a welter of detail about agents' incentives in imperfect markets.[9]

Many game theorists and other economists attended summer seminars at Stanford, with prominent game theorists being among the regulars, notably Robert Aumann and colleagues from Israel, where mathematically rigorous work was most advanced. The focus was essentially about how to formulate and analyze economic models in which all agents are strategic players, and whether such efforts would be useful in economic theory and applications. Essential roles were played by Kenneth Arrow, whose influential 1963 article on markets for medical care and insurance recognized these ingredients as intrinsic to the problem of organizing such markets, and Leonid Hurwicz, whose 1973 article showed the impediments posed by agents' private information and strategic behavior, and the necessity of taking account of them in economic analysis.

Hurwicz formulated the concept of a mechanism for implementing social choices, specified as a procedure that uses messages received from agents to select an outcome. The messages could be reports of privately known preferences or information, and the outcome could be an allocation of goods and/or selection of public projects. He invoked Nash equilibrium as a predictor of agents' strategic behavior, and more generally, emphasized the constraints imposed by incentive

[8] Early on, economists recognized that the apparatus of game theory enabled precise description of "who knows what when" and their available actions. Its solution concepts were problematic, but its descriptive power exceeded the usual tools of microeconomics. The *International Journal of Game Theory* began publication in 1971. By the late 1970s, game theory was widely adopted as a basic tool for modeling and analysis in theoretical microeconomics.

[9] In economics, the roles of private information and imperfect observability of actions (hidden information and hidden actions in Arrow's phrasing) were suddenly on display in Akerlof (1970) on markets for "lemons" and Mirrlees (1971) on optimal taxation, and somewhat later in Spence (1973) on signaling in labor markets, Rothschild and Stiglitz (1976) on insurance markets with adverse selection, and Mirrlees (1979) and Holmstrom (1977) on optimal contracting. All posited simple behavior on one side of the market and studied optimal strategies of the other side, a style that came to characterize information economics.

compatibility as key to unifying classical and game-theoretic analyses. Familiar market institutions, such as auctions and exchanges, provided abundant examples of mechanisms. Moreover, his perspective encompassed descriptions of existing mechanisms (and perhaps understanding why market designs of ancient vintage work well in many contexts), and designing efficiency-improving modifications—the genesis of market design.

After 1976, it was well-established that game theory offered potentially useful models and analytical tools for studies of strategic behavior, especially in contexts with imperfect observability of agents' information and actions. Its impact was initially in industrial organization, where game-theoretic results rebutted some earlier conclusions; in labor, where it offered richer theories of contracting; and in experimental and empirical studies, where detailed structural models supplanted reduced-form regressions. Its use grew steadily until it was widely taught to PhD students in economics (who needed such skills to read proliferating journal articles that relied on it), the arrival of excellent texts for economists, and a surge of young scholars who invoked strategic analysis in their research.

Roth: Cooperative games were studied in "coalitional form" models that specified what each coalition of players could achieve on its own. The most tractable model was the "characteristic function with transferable utility" (or "with side payments"), which modeled a game among a set $N = \{1, \ldots, n\}$ of players by specifying what numerical payoffs each coalition—that is, each subset S of N—could assure for its members. The assumption of "transferable utility" meant that each coalition was able to distribute the maximum sum of payoffs that it could achieve in any way that it wished among its members. Hence a game could be represented by a vector of real numbers, one for each coalition. The "characteristic function form" of a game was a function v on the subsets of N, that is, a pair (v, N) with v: $2^N \rightarrow R_+$ representing how much each coalition could achieve on its own. Outcomes of the game could then be represented as payoff vectors, one to each player in the game.

Von Neumann and Morgenstern (1944) defined how one payoff vector y could dominate another payoff vector x via a coalition S, if the coalition S, acting on its own, could guarantee each member i of S a payoff y_i greater than the corresponding payoff x_i at the outcome x. They defined a "solution" of the game to be a set of feasible payoff vectors of the game, none of which dominated another element in the solution, but at least one of whose elements dominated any element outside of the solution. It was easy to construct games with a multiplicity of solutions, but they conjectured that there would exist no game (v, N) for which no solution existed. This conjecture was eventually disproved (Lucas 1969), but even before that, von Neumann–Morgenstern solutions had not proved to be useful in understanding many games, and fell from use in economics.[10]

[10] They didn't fall from use entirely, however, before I wrote my PhD dissertation on generalizations of von Neumann–Morgenstern solutions that had better existence properties (Roth 1975, 1976). Much progress in game theory was made by exploring and identifying dead ends.

However, their idea of domination proved quite useful, in coalitional games of all sorts and not just of the form (v, N), and particular attention started to be paid to the set of undominated outcomes of a game, named the core of the game. The core could gracefully be generalized to games without side payments, in which the set of outcomes that each coalition could guarantee to its members might have to be described by a set that could not be summarized by a single number.[11]

In either case, it is sometimes useful to think of the core as the set of outcomes for which no "blocking coalitions" can form and produce an outcome that its members prefer. In some of the applications to labor markets, we focus on small blocking coalitions, consisting of just a single firm and worker, who, if not matched to each other when they would prefer to be, can form a "blocking pair." Matchings of firms to workers that have no blocking pairs are called pairwise stable matchings, and in some simple models these coincide with the core (and of course core outcomes have no blocking pairs, since they have no blocking coalitions of any size). Even in applications in which the set of stable matchings is bigger than the core, it is often the subset of outcomes of the most interest, because it is easier for small blocking coalitions to form than it is for large ones. I'll come back to this when speaking of the clearinghouse through which American doctors find their first jobs.[12]

Models of games without side payments, which also had a long history, gradually replaced games with side payments as the primary models for particular kinds of cooperative games. For example, exchange economies were modeled as having each player endowed with a vector of continuously divisible commodities, and coalitions of players were able to trade freely among themselves. This was a model without side payments, in that the set of outcomes a coalition could achieve on its own couldn't be described by a single number but was rather the set of allocations to members of the coalition that could be reached by trade within the coalition.

Although the core is empty for many games, it is non-empty for these exchange economies since the core contains the competitive allocations. This reinforces the idea that, when it is non-empty, the core can be interpreted as a model of the outcomes that would result from perfect competition. This idea is reinforced by

[11] A lot of creative effort went into generalizing the core in ways that would give a non-empty set for all games (v, N), such as the bargaining set (Aumann and Maschler 1964) and the kernel and nucleolus (Maschler 1992). A very different solution concept was the Shapley (1953) value, also well defined as a unique outcome of any game (v, N), which was meant to capture something like the expected utility of playing the game, in each of its positions (see Roth 1988 for a collection of articles on the Shapley value collected in honor of Shapley's 65th birthday). Despite heroic attempts, the generalizations of the bargaining set and Shapley value never caught on. Regarding the Shapley (1969) value for games without transferable utility, see my exchange with Aumann in Roth (1980, 1986) and Aumann (1985, 1986), all of which are reprinted in Aumann (2000) along with some other closely related papers.

[12] The term "blocking coalition" has become standard, despite the fact that it may not be the term that best expresses the manner in which these coalitions make outcomes outside of the core less likely. Shapley (1973) suggested that they be called "improving" coalitions, so that the core could be defined as the set of outcomes upon which no coalition can improve.

the observation that if the economy grows large in appropriate ways, then the core shrinks to become precisely the set of competitive allocations.

Models of cooperative games without side payments that became very important in my own work were the two-sided marriage model of Gale and Shapley (1962), in which individuals on opposite sides of the market could match to one another if they both agreed, and the exchange economy with indivisible goods of Shapley and Scarf (1974). Both papers demonstrated algorithms that, for any specification of the players' preferences, would produce an outcome (a matching of pairs, or a redistribution of initial endowments, respectively) in the core of the game. In this way, both papers showed constructively that, for any preferences, the core of the game was non-empty. (Gale and Shapley concentrated primarily on a model in which the core and the set of pairwise stable outcomes coincide.)

Despite the development of some important models and results, cooperative game theory was in decline by the late '70s, as more game theorists took the point of view that came to be called the "Nash program," which is that all games could be modeled strategically. The idea was that if binding agreements were possible, then how they were reached should be modeled in the extensive form, so that all games could/should be modeled strategically, as noncooperative games.[13] Of course, to model and analyze complex games is difficult, so one consequence of this approach was that games studied by game theorists would have strategy sets that could be generated by a small set of rules.

What Was Missing That Was Needed for Practical Market Design?

Wilson: For market design to reach practice, the missing ingredients were theoretical and experimental studies that gave some confidence to predictions about how design features affect performance. Early applications were partly guesswork, but with accumulated experience and an increasing trove of scholarly studies, fewer informed guesses are needed. Game theory has been the principal analytical tool because it enables detailed modeling of agents' information, incentives, and feasible strategies and provides predictions about equilibrium behavior and outcomes. Theoretical and experimental exercises rely on simplistic models but they clarify and test the basic concepts applied in practical work where invariably the situation is more complicated than can be modeled precisely.

Designs of auctions and matching markets evolved from the disparate branches of noncooperative and cooperative game theory. Agents' private information is the main consideration in auctions, and designs focus on procedures that elicit demands and yield good outcomes. The goal is to implement Walras and Hayek using Hurwicz's scheme. This is straightforward when bidders simply know their own private values for items, but more complicated when their information includes

[13] For example, the important *Game Theory* textbook of Fudenberg and Tirole (1991) contained no mention of the core, or of any other solution concepts from cooperative game theory.

estimates of unobserved factors that ex post will affect all their realized values (for example, in an auction of spectrum licenses, customers' ultimate demands for uses of spectrum), and then multiple rounds of bidding can enhance implicit revelation of bidders' estimates via their bids. Auction designs often take the set of bidders as a datum, but as Al describes below, a matching market presents the more formidable challenge of yielding an outcome so good it attracts participants.

Roth: A big missing part of cooperative game theory involved how some features of the game that could be expected to be private information, such as the preferences of the players, would become known. This concern actually fit in well with the Nash program of modeling games strategically: for example, if we asked participants in a game to reveal their preferences, then their strategies would include stating preferences different from their true preferences. Under what circumstances would a "revelation game" of this sort elicit the information we might wish to know?

Wilson: Hurwicz (1973) considered such revelation games and argued, using the example of an Edgeworth box, that they could not be viable unless "incentive compatibility" constraints were imposed on the mechanism.

Roth: A big missing part of noncooperative game theory was how we would know if a game produced "bad" outcomes that some participants might wish to circumvent by engaging in a larger game that might not be fully visible to the theorist. When I started to study labor markets, I saw that firms and workers had very large strategy sets that allowed them to approach each other in many ways, and at many times. For example, when professional organizations tried to organize job markets for new doctors, or new lawyers, they specified rules of engagement between applicants and employers, but for many years these rules didn't succeed in organizing those markets because there were incentives for applicants and employers to find creative ways to work around them, often reaching agreements well before the markets were officially supposed to begin (Roth and Xing 1994). Analyzing those markets under the assumption that everyone played by "the rules" would have yielded different outcomes than were observed, and indeed many observed outcomes in labor markets involved strategies that were either not imagined or explicitly forbidden under the official rules.

Mechanism design was in the spirit of studying fully known games—a game would be designed, in all its parts, that would specify all possible strategies, so the players would have no options outside of the game (except perhaps, not playing at all). That worked well when the designer could make players play the game: for example, a company or government that wanted to sell or buy something and could define the rules of the auction which those who wanted to transact must participate in.[14] But

[14] Even in these cases, Klemperer (2004) for example emphasizes that governments' auctions of spectrum licenses or Treasury bonds may allow strategies outside the formal rules, such as pre-auction mergers of firms to reduce competition in auctions, and trades in post-auction secondary markets.

lots of markets don't have this kind of compulsory power and must persuade users to participate.

What cooperative game theory ideas like the core allow us to do is to see whether the equilibrium of a game played by the rules of a particular marketplace is an outcome in the core of the larger game in which coalitions can find ways to act on their own outside of the marketplace we are modeling strategically. If the equilibrium outcome of the strategic model is not in the core of the coalitional model, then there are some coalitions (for example, of firms and workers) who might have incentives to try to get a better outcome. Even if we don't know all the strategies available to them, that's a clue that the rules may be subject to attack and evasion by the dissatisfied parties. I'll say more about the complementary roles played by "noncooperative" and "cooperative" models of the same economic environment as I talk about the clearinghouse designs that were ultimately successful in organizing the market for new doctors, and the periodic market failures that continue to afflict the market for new lawyers.[15]

This is why market designers started to ask whether the equilibrium behavior elicited by particular rules led to outcomes that were in the core of the game. The idea is that trying to promote rules that lead to outcomes outside the core—that is, outcomes that leave some coalitions getting less than they might be able to get by acting on their own—might give potential marketplace participants incentives to transact outside the marketplace. The core and related formulations of stability give us a way of saying something about the fact that participants have strategies outside of the marketplace, and that successful marketplaces will be those that don't give participants reason to go elsewhere.

The big lesson of market design is that marketplaces are small institutions in a big economic environment: participants have bigger strategy sets than you can see, and there are lots of players, not all of whom may even be active participants in the marketplace, but can influence it. So we needed a way to design mechanisms that had both good equilibrium properties for the rules we knew about, and good stability properties for the strategies we didn't know about.

Thus, the connection between coalitional and strategic models as they can be used in market design is not as models of different kinds of games, but as models of a given game at different levels of detail, used for complementary purposes. For parts of the game that we're designing, we use "noncooperative" strategic models to precisely specify actions available to players. For parts of the game that we don't have complete control over, we use "cooperative" coalitional models to tell us something about the incentives that agents and coalitions of agents may have to circumvent the rules. The idea of focusing on, say, pairwise stability in two-sided matching models

[15] In Roth (1991a), I wrote of the separation of cooperative and noncooperative game theory, saying that the less-detailed cooperative models, which try to represent a game without specifying all the rules, aspire to a spurious generality (because the omitted details matter), while the noncooperative, strategic models, which are analyzed as if they represented all the potential moves in a game, offer a spurious specificity when the game in question is a model of some observable situation (because we can seldom know all the potential moves).

is that if a pair of agents is eager to match with each other despite the fact that the rules of the marketplace mechanism prevent them from doing so, then maybe their strategy sets will be big enough to find a way to match with each other. (But if just one of them is interested in matching with the other, it may be difficult for the unhappy player to find a way to circumvent the marketplace and force a match with an unenthusiastic partner.)

For example, in the job market for new doctors, before a centralized clearinghouse was adopted, and in some of the markets for new lawyers still, candidates are often hired years before employment will begin, and before the official rules of the market allowed hiring to begin (Roth 1984; Roth and Xing 1994; Avery, Jolls, Posner, and Roth 2001, 2007). That is, firms and workers who were dissatisfied with the way the official marketplace functioned were able to circumvent it by signing contracts before it opened. This problem was effectively solved for the medical market by a clearinghouse that produces stable matchings (Roth and Peranson 1999).

Wilson: This perspective is analogous to one in the older literature on general equilibrium. A modern view might aim to determine whether a perfectly competitive market is a mechanism yielding an allocation that is efficient, or better yet, in the core. Because a competitive allocation is easily shown to be efficient and in the core, a theorem establishing existence of equilibrium prices and the resulting allocation is, in effect, an affirmative answer when competition is sufficient to justify traders' price-taking behavior in response to prevailing equilibrium prices. Such a theorem typically suppresses all detail about how the market is organized and how prices are established, summarizing it all in traders' budget constraints.

Roth: So general equilibrium theory shares with cooperative game theory the goal of identifying likely outcomes without focusing on all the details of how they are achieved.

Wilson: The focus on properties of the allocation began with Edgeworth's (1881) informal argument that the core shrinks to the competitive allocations as the market becomes more competitive (by replicating traders), established formally by Debreu and Scarf (1963), and culminated in Aumann's (1964) proof that the core consists only of the competitive allocations when the set of traders is a non-atomic measure space (so that no trader is large enough to have market power). These results led to the modern view that an ideal price-mediated perfectly competitive market might indeed be a mechanism largely immune to institutional details that yields a core allocation, but realistically the challenge in practice remains to design a mechanism that yields a core allocation. The focus on the core, and coalition stability more generally, stems from the prediction that the mechanism will miss some gains from trade if other opportunities attract away some potential participants. The design problem is most acute in those matching markets without transfer payments, but it is relevant whenever the mechanism's performance depends on attracting wide participation. Many of the most successful auction designs addressed contexts where participation was mandated by a monopolist seller, such as government auctions of spectrum licenses or a system operator's auctions of access to power transmission, but attracting wide participation is paramount in newer applications such as the design of trading platforms that operate in competition with other venues.

Roth: Another way in which market design involves a bigger economic environment than a narrowly defined mechanism design problem is that it may have to take account of players who are not intended to be, and who do not intend to be, participants in the market. In particular, some transactions, and the markets that serve them, are "repugnant" in the sense that some people would like to participate in them, while other people (who may not have any apparent connection to these transactions) think that they shouldn't be allowed (Roth 2007). But successful markets require a degree of social support, so these concerns need to be taken into account if a marketplace is to succeed. Widely held feelings of repugnance often make it necessary for a market designer to study and understand the moral, ethical, and esthetic opinions of members of the society in which the market might function, as well as their professional and social codes of conduct and courtesy.

Much of my work in facilitating kidney transplants through the design of exchange mechanisms can be viewed as arising from the widespread repugnance to, and laws against, the purchase of organs for transplants.[16] And some of my current work on expanding kidney exchange internationally, while gaining gratifying support in some quarters, is also meeting with a repugnance reaction in others, including concerns that it might expand black markets in poor countries.[17]

Note that it is also a market design task to think about how and whether particular kinds of markets can be effectively banned, since laws seeking to ban markets often inadvertently serve to design illegal black markets. Like other kinds of marketplace designs, legal bans on markets also occupy a place in a larger economic environment, and may be difficult to effectively enforce without wide social support, or if the markets in question are available in other jurisdictions. Markets and marketplaces that are legal in some places but banned in others include markets for prostitution, surrogacy, marijuana, etc.

Finally, in addition to requiring an expanded view of what strategies players may have access to, and which players may be involved, market design also has to take into account that players may fail to coordinate on equilibrium behavior. In this regard, experiments have played an important role in exploring the gap between what perfectly rational players might deduce, and what ordinarily competent

[16] For examples, see Roth, Sönmez, and Ünver (2004, 2005a, b), Roth, Sönmez, Ünver, Delmonico, and Saidman (2006), Rees et al. (2009), Leider and Roth (2010), and Ashlagi, Gilchrist, Roth, and Rees (2011a, b). Notice that this selection of papers is drawn equally from analyses appearing in economics journals and in medical journals, which reveals something about how practical market designs are developed.

[17] Rees, Dunn et al. (2017) offer a new proposal to expand kidney exchange internationally, Delmonico and Ascher (2017) express opposition, and Rees, Paloyo et al. (2017) and Roth et al. (2017) reply.

humans might find difficult,[18] particularly in the absence of common knowledge that all players were perfectly rational (see Roth 2016 on experiments specifically aimed at market design).

Why Were Auctions and Two-Sided Matching Markets Such Fertile Ground for Market Design? How and Why Did Auction Design Proceed Differently from the Design of Matching Markets?

Wilson: Centralized auctions and matching markets were fertile grounds for market design due to coincidence of several features. The key design element specifies rules of a game. The contexts are often sufficiently circumscribed that, if the design yields efficient (or better, core) outcomes, then one can expect agents to play that game rather than some larger game with myriad other possibilities. And the contexts usually justify assumptions of rational optimizing behavior rather than various behavioral possibilities. Moreover, the game is sufficiently simple that it can, to a limited extent, be modeled and analyzed, or in any case rough predictions of performance can be based on applications of basic concepts, simulations, experimental evidence, and prior experience.

This simplicity accounts also for the profusion in academic journals of scholarly studies of these kinds of markets, including theoretical, experimental, and empirical studies. But the methodologies employed for studies of auction and matching markets differ.

Auction studies usually rely on preferences represented as expectations of net monetary values, the mechanism translates static or dynamically adjusted bids into an allocation, the objective is an equilibrium allocation that is efficient or revenue maximizing for the seller, and beyond that objective, the design task often focuses on rules that suppress “gaming the system.” Except in Vickrey auctions, there is no attempt to elicit agents’ true preferences; instead, one elicits a willingness-to-pay that already “shades” the bid to exploit monopoly power derived from small numbers of bidders and their private information, often called informational rents. Efficiency is hard to assure in cases, like spectrum auctions, where agents want to acquire packages of complementary goods, so designs aim for approximate efficiency. Ex post efficiency is the actual goal, but this is tenuous due to agents’ private information or estimates about common-value components.[19]

In contrast, studies of prominent matching markets rely on ordinal preferences solely about one’s assigned partner(s), the mechanism translates directly reported preferences into recommended assignments, and the objective is a core

[18] See, for example, Roth and Erev (1995) for discussion of games in which players learn quickly to play equilibrium and others (such as the ultimatum game) in which learning may be very slow, and see Li (2017) for discussions of how strategy-proof mechanisms may not be transparent to participants.

[19] The modern state-of-the-art in auction design is presented superbly in the two books by Milgrom (2014, 2017).

allocation. This stronger criterion discourages matches outside the mechanism (as Al described above), and in simple cases, such as a "marriage market," a core allocation based on reported preferences is obtained via Gale and Shapley's deferred acceptance algorithm, or Shapley and Scarf's (1974) top trading cycles. Moreover, truthful reporting is a dominant strategy for the proposing side (Dubins and Friedman 1981; Roth 1982), so only the other side might gain from other strategies.

Roth: "Design" is a noun as well as a verb, and market design has its origins in the noun, in the study of the designs of existing marketplaces, and how different designs—different marketplace institutions, rules, and customs—can induce different strategies and produce different outcomes. Centralized marketplaces are a good place to start the study of market designs, because, by virtue of being centralized, a significant portion of their design may reside in well-codified rules and procedures that are easy to observe. For the same reason, when it becomes necessary to design new rules and procedures, the work involved in designing centralized marketplaces can have a very mechanism-design "look and feel," with well-defined kinds of messages communicated and processed in precisely specified ways that offer a concrete path to implementation in practice.

Auctions are centralized marketplaces in which the messages are bids, and the auction rules determine the form that bids take, how they are communicated, and how they determine the resulting payments and allocation of the items being auctioned. Because auctions are ancient tools of demand elicitation, practical knowledge about auctions began to be developed fairly early and game theory allowed auction theory to be formalized and extended as one of the early successes of the theory of mechanism design (for example, Vickrey 1961; Milgrom and Weber 1982). The view that auction rules can be designed was enhanced by Cassidy's (1967) survey of the vast variety of auctions used in practice, with differing incentives and performance.

In addition, if the goods being sold are available only from a single seller, then an auction satisfies the implicit assumption of mechanism design theory that purchasers must participate in the auction if they wish to buy. (For example, oil drilling or timber cutting licenses sold by the Department of the Interior, spectrum licenses sold by the Federal Communications Commission, and advertisements sold by Google connected to searches on their search engine are each sold by a single seller.) Thus, at least to a first approximation, the strategies that the auction designer makes available are the strategies that the bidders must use, and (some appropriate refinement of) strategic equilibrium among those strategies may be a good guide to designing the market and predicting the outcome.

Wilson: Of course, the seller should also abide by the rules, or have an incentive to do so to the extent observable by bidders. This criterion is implied by the definition of a credible mechanism proposed by Akbarpour and Li (2018); for an historical application, see Engelbrecht-Wiggans (1988).

Mechanism design theory usually imposes an "individual rationality" constraint that no agent is worse off from participating, but this constraint is very weak compared to the stronger requirement that agents prefer the centralized market to contracting outside the market, which is achieved by a mechanism yielding a core allocation. Electricity markets suggest another paradigm: actual energy flows must be determined by bids in the transmission operator's centralized market, but participants often contract bilaterally via long-term contracts or buy financial hedges pegged to the operator's real-time prices for energy and transmission. Decentralized hedging markets ameliorate price volatility in the operator's market, and each relies on the other to function well.

Roth: In contrast to auctions, labor markets, kidney transplants, and other matching markets start off very decentralized. Labor markets have many applicants and employers, and kidney transplants in the US are performed at hundreds of hospitals, each with considerable autonomy. So rather than being able to design a marketplace that all participants must use, the design of marketplaces for many matching markets involves finding designs that will entice users to try them, and satisfy them well enough that they will accept the outcomes and continue to come back to the marketplace. But when these designs lead to centralized clearinghouses, the marketplace itself nevertheless has considerable mechanism design flavor (as a stand-alone game) when one concentrates on the options available to participants within the marketplace. So these marketplaces were also a good starting point for market design.

The American marketplace for new doctors, the National Resident Matching Program, developed such a clearinghouse in the early 1950s in response to widespread market failures of the various decentralized market designs that had been employed in the first half of the 20th century. When I studied it in Roth (1984), I found that, by a process involving more than a little trial and error, the market had become organized since 1952 by a centralized clearinghouse in which candidates and employers submitted rank order lists of one another, and a centralized algorithm produced a suggested match. The algorithm that had been settled on turned out to be essentially equivalent to the hospital-proposing deferred acceptance algorithm studied a decade later by Gale and Shapley (1962). So on the one hand, this was a "mechanism" whose design could be studied, but on the other hand, one question that had to be answered was why this design had been enticing enough to attract the lion's share of the market. This was a pressing question when the marketplace needed to be redesigned in the mid-1990s. A critical fact, discovered by comparing successful and unsuccessful clearinghouses (Roth 1991b) and by experimentation (Kagel and Roth 2000), was that the stability of the resulting outcome was an important factor in its success. This was a clear example of the complementary uses of strategic and coalitional models in understanding the success of a marketplace.

Note how this reflects how game-theoretic ideas about the core and stable matchings have evolved as they have been confronted by the realities of market design. When we used to think of a game as a whole world, we often thought of the core as a model of what players would coordinate on, based on complete information about the game. So if players didn't know each other's preferences, blocking coalitions would be difficult to identify, and the core (that is, the set of outcomes for which there aren't any blocking coalitions) might not have much predictive power. But in a labor clearinghouse like the medical match, no one knows everyone else's preferences: it's a game of massively incomplete information. Yet, empirically, matching algorithms that produce unstable outcomes fail. Because the clearinghouse is part of a larger economic environment, it's no mystery how this can happen. If I am matched to my third-choice residency program, I only have to make two phone calls to find out if I am part of a blocking pair—that is, if one of the residency programs I prefer might also prefer me to one of the doctors it has been matched with. If in previous years many matches were found that way, then this year people will make those phone calls too, and the clearinghouse will fail to organize the market because residency programs and individual doctors will make deals on their own, and not accept those produced by the clearinghouse. But a stable matching algorithm will be robust to phone calls, since there aren't any blocking pairs to find. Those phone calls aren't part of the description of the clearinghouse, they are part of the larger economic environment in which the clearinghouse is just a small marketplace.

In contrast to the market for new doctors, the market for judicial clerks uses rules that have been regularly designed and subsequently abandoned by judges themselves, and have yet to find a market design that entices judges to participate according to the rules (Avery et al. 2001, 2007, and my blog post at https://marketdesigner.blogspot.com/search/label/clerks).

Wilson: I was fascinated by the Roth and Xing (1994) article describing many markets that work imperfectly because they fail to deter contracting outside the market including prior contracting (to snag a good partner before others do), as well as backup plans to try again in a decentralized aftermarket among those not satisfied by the recommended assignment.

A novel feature of some matching markets is assignments recommended to agents, rather than binding contracts, and most agents voluntarily accept their recommended matches: the deferred acceptance algorithm assures that no two agents prefer each other to their assigned partners. Some simple auctions have this property, and there are designs that aim for it, but generally the prevalence of private information precludes assurance of a core allocation based on revealed preferences and information.[20] Government agencies typically use auctions designed to

[20] Day and Milgrom (2008) derive key properties of a core-selecting auction when one exists, and relate this to the stable matching literature. Kelso and Crawford (1982) consider an auction that closely resembles the deferred acceptance algorithm, in a labor market context in which the auction chooses both a matching and the associated market-clearing doubly personalized wages (that is, wages for each firm-worker pair).

yield an efficient allocation, but alternative designs forego efficiency to maximize the seller's expected revenue.

Roth: How about in electricity markets?

Wilson: Just as you said before, participants in electricity markets have big strategy sets, and there are interested parties who aren't participants in the market. Firms can bid in the spectrum auctions of the Federal Communications Commission, buy licenses in secondary markets (if the FCC approves the transfer), or rent spectrum from those with licenses. In wholesale electricity markets, firms must bid in the system operator's daily and hourly energy auctions to get power scheduled for transmission, and the operator also runs various auxiliary auction markets for transmission rights (which hedge transmission charges) and reserves of capacity available in various time frames. But these are minor parts of the overall market because most power is contracted long term. A typical bilateral contract for delivery has an agreed price, and the parties settle the difference between their price and the operator's price.

There are also financial markets for financial instruments that hedge against volatility of the operator's prices. On the supply side there are further markets for fuel, especially long-term contracts for natural gas that ensure priority when supplies are tight. And on the demand side there are markets for demand reduction by firms who can curtail or interrupt power usage when prices are high.

Thus, no one of these auctions is isolated; rather, each operates within a loosely coordinated system of related markets. This system need not yield an outcome in the core like in a matching market because no analog of the deferred acceptance algorithm has been found, so it relies on competitive pressures to ensure efficient outcomes—and part of the design task is to promote vigorous competition. Besides participants, there are other important actors who affect the system design, most importantly the Federal Energy Regulatory Commission that prescribes standards for system operators, each state's public utility commission, which regulates retail distribution (and in some cases appoints the board of the system operator), and federal agencies that regulate commodity trading and financial markets for commodity futures contracts.

Roth: And are aspects of electricity sales repugnant?

Wilson: The repugnance factor can be traced in the history of the restructuring of electricity markets. Some view electricity as a necessary service that is best provided by vertically integrated utilities. Prior to restructuring, this was implemented in each state by tight regulation that set retail service standards and prices and in return provided utilities with an assured rate of return on capital. In most states, this regime dissolved because high prices for retail service were attributed to distorted incentives, resulting in excessive capital intensity manifest in massive power plants, and monopolization of transmission that disadvantaged independent power producers.

After political battles, the industry was restructured by most states requiring utilities to divest their generation assets, and federal requirements for open access to transmission via auction markets conducted by independent system operators. Not all states restructured. But even in those that did, lesser battles continue, represented for example by newly formed municipal or cooperative power distribution companies that opt out of utilities' retail services by buying supplies directly from system operators' markets. Basically, market design for electricity markets aspires to the ideal service that vertically integrated utilities were intended to achieve, but now implemented in an open-access decentralized system with stronger incentives.

How Was Your Work Influenced by Your Teachers, Colleagues, and Students?

Wilson: The interests of my advisor Howard Raiffa had turned to statistical decision theory, but I retained interest in game theory, motivated by a colleague's study of investment banking syndicates formed to bid for corporate bonds. After some early consulting on corporate strategies, my practical work on market design began in consulting with the US Department of Interior's section on oil exploration led by Darius Gaskins. He hired me because he had concluded that game theory was needed to analyze their auctions of licenses. There were aspects of auction design, but I focused mainly on algorithms for bidding strategies based on models that included adverse selection—aka "the winner's curse"—and methods for ex post analyses of auction performance. This was also a focus in the late 1970s of my consulting with oil companies, especially with George Harwell at Natomas, but in these cases my primary job was to help them understand the effects of adverse selection. I learned a lot from being inside a company, watching how bids were derived by interpreting geological data to obtain (vaguely probabilistic) estimates that were then combined with data about costs and predictions of future oil prices. Equally educational was to hear insistence on finding the minimum bid that would win. Some dismissed adverse selection as hokum, but a few old sages insisted it was real and claimed they had survived in a fiercely competitive industry by using rules of thumb that severely cut engineers' estimates to be on the safe side.

Roth: The adverse selection that Bob is referring to, the "winner's curse," is that when each bidder gets an estimate of how much oil is under the ground at a given site, the bidder who has the highest estimate is very likely to have an overestimate. And the more bidders there are, the bigger the amount of the overestimation. Wilson (1977) introduced the model of common-value auctions (sometimes called the "mineral rights model"). The model and its equilibrium initiated a large body of theoretical, experimental, and applied work. One important insight from this model is that winning an auction contains "bad" news, since it implies in equilibrium that the winner's estimate is the highest. In equilibrium, rational bidders fully account for this, but the paper raises the empirical question of the extent to which actual bidders are able to fully discount for the fact that, if they win the auction, they likely overestimated the value of winning. Thus, Wilson's work initiated a new research program on the winner's curse, involving systematic overbidding compared to equilibrium, sometimes involving losses to the winning bidder.[21] The private-value model of Vickrey (1961) and the common-value model of Wilson (1977) together form the basis of much of modern auction theory and practice, since most auctions have elements of both private and common value.

Wilson: Another formative experience was at the Electric Power Research Institute in a section led by Steven Peck and Hung-po Chao, working with them and my co-consultant Shmuel Oren. It focused initially on innovative contracts for utilities' retail services, but over the ensuing 40 years its scope expanded to include all the issues posed by fundamental restructuring of the industry, and most relevant here, the design of centralized markets for energy, transmission, and reserves.

I was deeply affected in the early 1990s by working with Paul Milgrom on design of the FCC spectrum auctions. I marveled at his insights and creativity in constructing rules for a "simultaneous ascending auction" that would have good prospects of yielding an approximately efficient outcome in an environment afflicted with strong complementarities, dispersed private information about market fundamentals, and substantial market power. And we were greatly influenced by Evan Kwerel, who was the main protagonist at the FCC seeking innovative auction designs for allocating spectrum licenses.

Roth: I was also much influenced by Paul when we developed and co-taught what may have been the first courses in market design, in 2000 and again in 2001 when he was on leave at Harvard and MIT.

Wilson: The strongest influences on my work in game theory came from Robert Aumann's articles and lectures, and later, collaborations with David Kreps and Srihari Govindan. Even after I pursued market design, I continued my interest in foundations, seeking the full implications of rationality in multiperson interactions. I was also deeply influenced by superb PhD students, of which some directly affected my work in game theory and ultimately market design. Before 1980 they included Armando Ortega-Reichert, Robert Rosenthal, Alvin Roth, Jean-Pierre Ponssard, Claude d'Aspremont, Paul Milgrom, and Bengt Holmstrom. Later, the chief influence was Peter Cramton.

[21] In an early use of experimental economics to elucidate this issue, Bob invented the now famous "jar of coins" experiment, in which the value of the coins in a jar is auctioned off to the highest bidder. If every bidder forms his own estimate of the value of the coins (for example, of how many coins are in the jar), then the high bidder almost invariably has an overestimate, and, failing to account for this, bids more than the jar is worth. Used as a demonstration, this helps convince skeptics that the winners' curse is real (for a fuller account, see Roth 2016).

Roth: I would have had a very brief academic career if not rescued by Bob after flunking my qualifying exams. As I've learned also from my students, the teacher–student relationship can be one of life's big ones, although less appreciated than the other big relationships.[22] I've learned from (and designed with) students, postdocs, research fellows, and colleagues, some of whom fall in more than one category (and who I refrain from enumerating here only because the list has such ill-defined boundaries and includes so many of my students and coauthors that I would inevitably err by omission). I've also profited enormously from collaborating with practitioners. I think that virtually all of the market designs I've been involved in that were adopted and successfully implemented benefited from collaboration with someone involved in the market who became the champion of the new design.

How Have Other Domains of Market Design Developed?

Roth: I'd like to highlight two other domains I think foreshadow further ways market design is developing into a robust part of economics. The first is the design of school choice systems, which in its origins closely resembles the stable matching deployed in the clearinghouse marketplaces for doctors (Abdulkadiroğlu and Sönmez 2003; Abdulkadiroğlu, Pathak, and Roth 2005, 2009; Abdulkadiroğlu, Pathak, Roth, and Sönmez 2005). School choice has now opened a new window for the empirical study of schools, as econometric tools take advantage of the particular elements of the market design to measure not only the effects of the new designs on how students are assigned to schools, but also the importance to students of being well matched to a school (for example, Abdulkadiroğlu, Agarwal, and Pathak 2017; Abdulkadiroğlu, Angrist, Narita, and Pathak 2017; Agarwal and Somaini 2018).[23] I think of this as a kind of third generation of market design, since those of us initially involved in design were game theorists, who of necessity became engineers to help new designs be implemented and maintained, and we are now seeing those designs and their outcomes subjected to, and enabling, sophisticated empirical scrutiny by applied economists able to develop new econometric tools informed by the details of the markets' designs.

Another area of market design involves decentralized markets. Most markets are decentralized at least to some degree, and many almost entirely. Even markets

[22] A brief account of our subsequent teacher–student interactions, along with some other remembrances related to the present essay, is included in my intellectual autobiography at the Nobel Prize website: https://www.nobelprize.org/prizes/economics/2012/roth/auto-biography/. The rabbinic literature does not overlook teacher–student relations. In the Talmud, for example, one is enjoined: "Provide for yourself a teacher and get yourself a friend …" See my related blog post for more on this: https://marketdesigner.blogspot.com/2013/06/notes-on-teachers-and-students-from.html. The martial arts also value teacher–student relations, and I benefited from that too, as I describe at https://marketdesigner.blogspot.com/2013/06/honorary-7th-dan-black-belt-in-jka.html.

[23] Agarwal (2015) similarly uses econometric tools that leverage the stable matchings arising from the resident match to do a demand analysis of the market for different residency programs.

that employ centralized marketplaces may be preceded or followed by decentralized interaction. For example, the market for new academic economists has a somewhat centralized marketplace for interviews, preceded by decentralized applications and followed by decentralized campus visits, offers, acceptances, and rejections.[24]

Designers who wish to introduce more centralized marketplaces into existing markets need to understand how this will interact with pre-existing decentralized markets. In this respect, changing the presumptions about how (decentralized) offers were made, and until when they should remain open, helped pave the way for a centralized clearinghouse for gastroenterologists to become successful (Niederle and Roth 2003, 2009; Niederle, Proctor, and Roth 2006, 2008; McKinney, Niederle, and Roth 2005). PhD programs similarly are expected to leave offers of admission open until April 15 (Roth and Xing 1994). Design in decentralized markets may involve helping to form expectations and customs to promote and to guide decentralized transactions among market participants, rather than crafting precise rules and algorithms that can be rendered in computer code.

Wilson: A remarkable application is the conversion in the United States of spectrum from television broadcast to smartphones (Leyton-Brown, Milgrom, and Segal 2017). Ultimately this was done by one auction that bought spectrum from broadcasters and another that re-sold it to phone companies. The major complication was development of algorithms to reassign retained broadcast rights to new spectrum so as to avoid electromagnetic interference among broadcasting stations.

What's Next for Game Theory and Market Design?

Wilson: The ongoing computerization of marketplaces will continue to make market design a multidisciplinary endeavor, which already occupies computer scientists as well as economists.[25] And economic engineering more broadly—"design economics"—will likely continue to grow in its ability to help structure contracts, firms, and organizations and collaborations of all sorts.

[24] There have been some modest further efforts to aid coordination through centralized signaling before interviews and a scramble afterwards (Coles, Cawley, Levine, Niederle, Roth, and Siegfried 2010).

[25] For examples of some collaborations between economics and computer science, see Anderson, Ashlagi, Gamarnik, and Roth (2015) and Leyton-Brown, Milgrom, and Segal (2017). As computer and communication technologies increase the proportion of transactions conducted over highly automated and tightly coordinated platforms, we foresee the design of these platforms as a major area for market design involving joint efforts among economists, computer scientists, and software engineers relying on developments in "algorithmic game theory." Already, parameters of some markets are adjusted automatically by machine learning algorithms, and we are entering a time when market participants themselves may be designed—to some extent this has already happened in the realm of high-speed algorithmic trading of securities (as discussed in Budish, Cramton, and Shim 2015).

Roth: Smartphones have put marketplaces in our pockets, and as computerized marketplaces become ever more ubiquitous, we will also generate data trails that will continue to extend the reach of markets, socially and personally. We will learn more about privacy and fairness, and there will be new opportunities, some of which will come to be seen as repugnant, for which new market mechanisms, rules, customs, and regulations will have to be designed.

What's Your Last Word for Now?

Wilson: We've learned that maximizing gains from trade is more about participants' information and incentives than intersecting demand and supply curves. So concepts from game theory have been useful guides in efforts to improve the performance of trading platforms. But scholarly theorizing is minor compared to hands-on engineering using knowledge of an industry's technology and practices, and familiarity with participants' concerns is necessary if one is to help them obtain better outcomes overall. Deep involvement discovers key features unanticipated by abstract views of markets. I foresee more economists improving the allocation of scarce resources rather than (just) studying it.

Roth: Market design is an outward-facing part of economics, so designers have to be good listeners, prepared to learn from everyone. And learning from markets and their participants is a great driver of economic theory. One of my favorite quotes from Wilson (1993) is "for the theorist, the problems encountered by practitioners provide a wealth of topics."

■ *This paper benefited from comments by Vic Fuchs, Dave Kreps, and Ben Roth, none of whom endorse any of the views expressed here.*

References

Abdulkadiroğlu, Atila, Nikhil Agarwal, and Parag A. Pathak. 2017. "The Welfare Effects of Coordinated Assignment: Evidence from the New York City High School Match." *American Economic Review* 107(12): 3635–89.

Abdulkadiroğlu, Atila, Joshua D. Angrist, Yusuke Narita, Parag A. Pathak. 2017. "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation." *Econometrica* 85(5): 1373–1432.

Abdulkadiroğlu, Atila, Parag A. Pathak, and Alvin E. Roth. 2005. "The New York City High School Match." *American Economic Review* 95(2): 364–67.

Abdulkadiroğlu, Atila, Parag A. Pathak, and Alvin E. Roth. 2009. "Strategy-Proofness versus Efficiency in Matching with Indifferences: Redesigning the NYC High School Match." *American Economic Review* 99(5): 1954–78.

Abdulkadiroğlu, Atila, Parag A. Pathak, Alvin E. Roth, and Tayfun Sönmez. 2005. "The Boston Public School Match." *American Economic Review*

95(2): 368–71.

Abdulkadiroğlu, Atila, and Tayfun Sönmez. 2003. "School Choice: A Mechanism Design Approach." *American Economic Review* 93(3): 729–47.

Agarwal, Nikhil. 2015. "An Empirical Model of the Medical Match." *American Economic Review* 105(7): 1939–78.

Agarwal, Nikhil, and Paulo Somaini. 2018. "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism." *Econometrica* 86(2): 391–444.

Akbarpour, Mohammed, and Shengwu Li. 2018. "Credible Mechanisms." Stanford University. Available at SSRN: https://ssrn.com/abstract=3033208.

Akerlof, George A. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84(3): 488–500.

Anderson, Ross, Itai Ashlagi, David Gamarnik, and Alvin E. Roth. 2015. "Finding Long Chains in Kidney Exchange Using the Traveling Salesman Problem." *PNAS* 112(3): 663–68.

Arrow, Kenneth J. 1963. "Uncertainty and the Welfare Economics of Medical Care." *American Economic Review* 53(5): 941–73.

Ashlagi, Itai, Duncan S. Gilchrist, Alvin E. Roth, and Michael A. Rees. 2011a. "Nonsimultaneous Chains and Dominos in Kidney-Paired Donation—Revisited." *American Journal of Transplantation* 11(5): 984–94.

Ashlagi, Itai, Duncan S. Gilchrist, Alvin E. Roth, and Michael A. Rees. 2011b. "NEAD Chains in Transplantation." *American Journal of Transplantation* 11(12): 2780–81.

Aumann, Robert J. 1964. "Markets with a Continuum of Traders." *Econometrica* 32(1–2): 39–50.

Aumann, Robert J. 1976. "Agreeing to Disagree." *Annals of Statistics* 4(6): 1236–39.

Aumann, Robert J. 1985. "On the Non-Transferable Utility Value: A Comment on the Roth–Shafer Examples." *Econometrica* 53(3): 667–78.

Aumann, Robert J. 1986. "On the Non-Transferable Utility Value: Rejoinder." *Econometrica* 54(4): 985–89.

Aumann, Robert J. 2000. *Collected Papers,* vol. 2. Cambridge, MA: MIT Press.

Aumann, Robert J., and Michael Maschler. 1964. "The Bargaining Set for Cooperative Games." Chap. 21 in *Advances in Game Theory,* edited by M. Dresher, L. S. Shapley, and A. W. Tucker. Annals of Mathematics Studies 52. Princeton University Press.

Avery, Christopher, Christine Jolls, Richard A. Posner, and Alvin E. Roth. 2001. "The Market for Federal Judicial Law Clerks." *University of Chicago Law Review* 68(3): 793–902.

Avery, Christopher, Christine Jolls, Richard A. Posner, and Alvin E. Roth. 2007. "The New Market for Federal Judicial Law Clerks." *University of Chicago Law Review* 74(2): 447–86.

Borel, Emile. 1921. "La theorie du jeu et les equations integrales anoyau symmetrique gauche." *Comptes Rendus de l'Academie des Sciences* 173: 1304–08.

Borel, Emile. 1953. "The Theory of Play and Integral Equations with Skew Symmetric Kernels." [Translation by L. J. Savage of Borel (1921)] *Econometrica* 21(1): 97–100.

Budish, Eric, Peter Cramton, and John Shim. 2015. "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response." *Quarterly Journal of Economics* 130(4): 1547–1621.

Cassidy, Ralph, Jr. 1967. *Auctions and Auctioneering.* University of California Press.

Coles, Peter, John Cawley, Phillip B. Levine, Muriel Niederle, Alvin E. Roth, and John J. Siegfried. 2010. "The Job Market for New Economists: A Market Design Perspective." *Journal of Economic Perspectives* 24(4): 187–206.

Copeland, Arthur H. 1945. "Review: John von Neumann and Oskar Morgenstern, Theory of Games and Economic Behavior." *Bulletin of the American Mathematical Society* 51(7): 498–504.

Day, Robert, and Paul Milgrom. 2008. "Core-Selecting Package Auctions." *International Journal of Game Theory* 36(3–4): 393–407.

Debreu, Gerard, and Herbert Scarf. 1963. "A Limit Theorem on the Core of an Economy." *International Economic Review* 4(3): 235–46.

Delmonico, Francis L., and Nancy L. Ascher. 2017. Opposition to Irresponsible Global Kidney Exchange. *American Journal of Transplantation* 17(10): 2745–46.

Dubins, Lester E., and Daniel A. Friedman. 1981. "Machiavelli and the Gale–Shapley Algorithm." *American Mathematical Monthly* 88(7): 485–94.

Edgeworth, Francis Ysidro. 1881. *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences.*

Engelbrecht-Wiggans, Richard. 1988. "An Example of Auction Design: A Theoretical Basis for 19th Century Modifications to the Port of New York Imported Goods Market." Unpublished paper, University of Illinois.

Fudenberg, Drew, and Jean Tirole. 1991. *Game Theory.* MIT Press.

Gale, David, and Lloyd S. Shapley. 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly* 69(1): 9–15.

Govindan, Srihari, and Robert Wilson. 2012.

"Axiomatic Equilibrium Selection for Generic Two-Player Games." *Econometrica* 80(4): 1639–99.

Harsanyi, John C. 1967–68. "Games with Incomplete Information Played by 'Bayesian' Players." Parts I–III. *Management Science*, 14 (3, 5, 7): 159–82, 320–34, 486–502.

Hayek, F. A. 1945. "The Use of Knowledge in Society." *American Economic Review* 35(4): 519–30.

Holmstrom, Bengt. 1977. *On Incentives and Control in Organizations.* PhD dissertation, Stanford University.

Hurwicz, Leonid. 1945. "The Theory of Economic Behavior." *American Economic Review* 35(5): 909–25.

Hurwicz, Leonid. 1973. "The Design of Mechanisms for Resource Allocation." *American Economic Review* 63(2): 1–30.

Kagel, John H., and Alvin E. Roth. 2000. "The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment." *Quarterly Journal of Economics* 115(1): 201–35.

Karlin, Samuel. 1959. *Mathematical Methods and Theory in Games, Programming, and Economics*, 2 vols. New York: Wiley.

Kelso, Alexander S., and Vincent P. Crawford. 1982. "Job Matching, Coalition Formation, and Gross Substitutes." *Econometrica* 50(6): 1483–1504.

Klemperer, Paul. 2004. *Auctions: Theory and Practice.* Princeton University Press.

Kreps, David M., and Robert Wilson. 1982. "Sequential Equilibria." *Econometrica* 50(4): 863–94.

Kuhn, Harold W. 1950. "Extensive Games." *PNAS* 36(10): 570–76.

Kuhn, Harold W. 1953. "Extensive Games and the Problem of Information." Chap. 11 in *Contributions to the Theory of Games*, vol. 1, edited by Harold W. Kuhn and Albert W. Tucker. Princeton University Press.

Leider, Stephen, and Alvin E. Roth. 2010. "Kidneys for Sale: Who Disapproves, and Why?" *American Journal of Transplantation* 10(5): 1221–27.

Lemke, Carlton E., and J. T. Howson, Jr. 1964. "Equilibrium Points of Bimatrix Games." *Journal of the Society for Industrial and Applied Mathematics* 12(2): 413–23.

Lewis, David. 1969. *Convention: A Philosophical Study.* Cambridge, MA: Harvard University Press.

Leyton-Brown, Kevin, Paul Milgrom, and Ilya Segal. 2017. "Economics and Computer Science of a Radio Spectrum Reallocation." *PNAS* 114(28): 7202–09.

Li, Shengwu. 2017. "Obviously Strategy-Proof Mechanisms." *American Economic Review* 107(11): 3257–87.

Lucas, William F. 1969. "The Proof that a Game May Not Have a Solution." *Transactions of the American Mathematical Society* 137: 219–29.

Luce, R. Duncan, and Howard Raiffa. 1957. *Games and Decisions: Introduction and Critical Survey.* New York: John Wiley and Sons.

Marschak, Jacob. 1946. "Neumann's and Morgenstern's New Approach to Static Economics." *Journal of Political Economy* 54(2): 97–115.

Maschler, Michael. 1992. "The Bargaining Set, Kernel, and Nucleolus." Chap. 18 in *Handbook of Game Theory with Economic Applications,* vol. 1, edited by Robert J. Aumann and Sergiu Hart. Elsevier.

McKinney, C. Nicholas, Muriel Niederle, and Alvin E. Roth. 2005. "The Collapse of a Medical Labor Clearinghouse (and Why Such Failures Are Rare)." *American Economic Review* 95(3): 878–89.

Mertens, Jean-Francois. 1989. "Stable Equilibria—A Reformulation: Part I. Definition and Basic Properties." *Mathematics of Operations Research* 14(4): 575–625.

Milgrom, Paul. 2004. *Putting Auction Theory to Work.* Cambridge, MA: Cambridge University Press.

Milgrom, Paul. 2017. *Discovering Prices: Auction Design in Markets with Complex Constraints.* New York: Columbia University Press.

Milgrom, Paul, and Joshua Mollner. 2018. "Equilibrium Selection in Auctions and High Stakes Games." *Econometrica* 86(1): 219–61.

Milgrom, Paul R., and Robert J. Weber. 1982. "A Theory of Auctions and Competitive Bidding." *Econometrica* 50(5): 1089–1122.

Mirrlees, James A. 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies* 38(2): 175–208.

Mirrlees, James A. 1979. "The Implications of Moral Hazard for Optimal Insurance." Paper presented at conference in honor of Karl Borch, Bergen, Norway, 1979.

Myerson, Roger B. 1999. "Nash Equilibrium and the History of Economic Theory." *Journal of Economic Literature* 37(3): 1067–82.

Myerson, Roger B., and Jörgen Weibull. 2015. "Tenable Strategy Blocks and Settled Equilibria." *Econometrica* 83(3): 943–76.

Nash, John F., Jr. 1950a. "Equilibrium Points in *n*-Person Games." *PNAS* 36(1): 48–49.

Nash, John F., Jr. 1950b. "The Bargaining Problem." *Econometrica* 18(2): 155–62.

Nash, John F., Jr. 1951. "Noncooperative Games." *Annals of Mathematics* 54: 289–95.

Niederle, Muriel, Deborah D. Proctor, and Alvin E. Roth. 2006. "What Will Be Needed for the New GI Fellowship Match to Succeed?" *Gastroenterology* 130(1): 218–24.

Niederle, Muriel, Deborah D. Proctor, and Alvin E. Roth. 2008. "The Gastroenterology Fellowship Match—The First Two Years." *Gastroenterology* 135(2): 344–46.

Niederle, Muriel, and Alvin E. Roth. 2003.

"Unraveling Reduces Mobility in a Labor Market: Gastroenterology With and Without a Centralized Match." *Journal of Political Economy* 111(6): 1342–52.

Niederle, Muriel, and Alvin E. Roth. 2009. "Market Culture: How Rules Governing Exploding Offers Affect Market Performance." *American Economic Journal: Microeconomics* 1(2): 199–219.

Ortega-Reichert, Armando. 1969. *Models for Competitive Bidding Under Uncertainty*. PhD dissertation, Stanford University.

Owen, Guillermo. 1968. *Game Theory*. Academic Press.

Rees, Michael A., Ty B. Dunn, Christian S. Kuhr, Christopher L. Marsh, Jeffrey Rogers, Susan E. Rees, Alejandra Cicero, Laurie J. Reece, Alvin E. Roth, Obi Ekwenna, David E. Fumo, Kimberly D. Krawiec, Jonathan E. Kopke, Samay Jain, Miguel Tan, and Siegfredo R. Paloyo. 2017. "Kidney Exchange to Overcome Financial Barriers to Kidney Transplantation." *American Journal of Transplantation* 17(3): 782–90.

Rees, Michael A., Jonathan E. Kopke, Ronald P. Pelletier, Dorry L. Segev, Matthew E. Rutter, Alfredo J. Fabrega, Jeffrey Rogers, Oleh G. Pankewycz, Janet Hiller, Alvin E. Roth, Tuomas Sandholm, Utku Ünver, and Robert A. Montgomery. 2009. "A Non-Simultaneous Extended Altruistic Donor Chain." *New England Journal of Medicine* 360(11): 1096–1101.

Rees, Michael A, Siegfredo R. Paloyo, Alvin E. Roth, Kimberly D. Krawiec, Obi Ekwenna, Christopher L. Marsh, A. J. Wenig, Ty B. Dunn. 2017. "Global Kidney Exchange: Financially Incompatible Pairs Are Not Transplantable Compatible Pairs." *American Journal of Transplantation* 17(10): 2743–44.

Roth, Alvin E. 1975. "A Lattice Fixed-Point Theorem with Constraints." *Bulletin of the American Mathematical Society* 81(1): 136–38.

Roth, Alvin E. 1976. "Subsolutions and the Supercore of Cooperative Games." *Mathematics of Operations Research* 1(1): 43–49.

Roth, Alvin E. 1980. "Values for Games without Sidepayments: Some Difficulties with Current Concepts." *Econometrica* 48(2): 457–65.

Roth, Alvin E. 1982. "The Economics of Matching: Stability and Incentives." *Mathematics of Operations Research* 7(4): 617–28.

Roth, Alvin E. 1984. "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory." *Journal of Political Economy* 92(6): 991–1016.

Roth, Alvin E. 1986. "On the Non-Transferable Utility Value: A Reply to Aumann." *Econometrica* 54(4): 981–84.

Roth, Alvin E., ed. 1988. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press.

Roth, Alvin E. 1991a. "Game Theory as a Part of Empirical Economics." *Economic Journal* 101: 107–114.

Roth, Alvin E. 1991b. "A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom." *American Economic Review* 81(3): 415–40.

Roth, Alvin E. 2007. "Repugnance as a Constraint on Markets." *Journal of Economic Perspectives* 21(3): 37–58.

Roth, Alvin E. 2016. "Experiments in Market Design." Chap. 5 in *Handbook of Experimental Economics,* vol. 2, edited by John H. Kagel and Alvin E. Roth. Princeton University Press.

Roth, Alvin E., and Ido Erev. 1995. "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term." *Games and Economic Behavior* 8: 164–212.

Roth, Alvin E., Kimberly D. Krawiec, Siegfredo Paloyo, Obi Ekwenna, Christopher L. Marsh, Alexandria J. Wenig, Ty B. Dunn, and Michael A. Rees. 2017. "People Should Not Be Banned from Transplantation Only Because of Their Country of Origin." *American Journal of Transplantation* 17(10): 2747–48.

Roth, Alvin E., and Elliott Peranson. 1999. "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design." *American Economic Review* 89(4): 748–80.

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver. 2004. "Kidney Exchange." *Quarterly Journal of Economics* 119(2): 457–88.

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver. 2005a. "A Kidney Exchange Clearinghouse in New England." *American Economic Review* 95(2): 376–80.

Roth, Alvin E., Tayfun Sönmez and M. Utku Ünver. 2005b. "Pairwise Kidney Exchange." *Journal of Economic Theory* 125(2): 151–188.

Roth, Alvin E., Tayfun Sönmez, M. Utku Ünver, Francis L. Delmonico, and Susan L. Saidman. 2006. "Utilizing List Exchange and Nondirected Donation through 'Chain' Paired Kidney Donations." *American Journal of Transplantation* 6(11): 2694–2705.

Roth, Alvin E., and Xiaolin Xing. 1994. "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions." *American Economic Review* 84(4): 992–1044.

Rothschild, Michael, and Joseph Stiglitz. 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Quarterly Journal of Economics* 90(4): 629–49.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Harvard University Press.

Selten, Reinhard. 1965. "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit." *Zeitschrift für die gesamte Staatswissenschaft*

121: 301–29, 667–89.

Selten, Reinhard. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4(1): 25–55.

Shapley, Lloyd S. 1953. "A Value for *n*-Person Games." Chap. 17 in *Contributions to the Theory of Games*, vol. 2, edited by Harold W. Kuhn and Albert W. Tucker. Princeton University Press.

Shapley, Lloyd S. 1969. "Utility Comparison and the Theory of Games." In *La Décision: Agrégation et Dynamique des Ordres de Préférence,* pp. 251–263. Paris: Editions du CNRS.

Shapley, Lloyd S. 1973. "Let's Block 'Block.'" *Econometrica* 41(6): 1201–02.

Shapley, Lloyd, and Herbert Scarf. 1974. "On Cores and Indivisibility." *Journal of Mathematical Economics* 1(1): 23–37.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87(3): 355–74.

Vickrey, William. 1961. "Counterspeculation, Auctions, and Competitive Sealed Tenders." *Journal of Finance* 16(1): 8–37.

von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior.* (2nd edition, 1947; 3rd edition, 1953). Princeton University Press.

Wald, Abraham. 1947. Review of "Theory of Games and Economic Behavior" by John v. Neumann and Oskar Morgenstern. *Review of Economics and Statistics* 29(1): 47–52.

Wilson, Robert B. 1977. "A Bidding Model of Perfect Competition." *Review of Economic Studies* 44(3) 511–18.

Wilson, Robert B. 1987. "Game-Theoretic Analyses of Trading Processes." In *Advances in Economic Theory: Fifth World Congress*, edited by Truman F. Bewley, pp. 33–77. Cambridge University Press.

Wilson, Robert B. 1993. *Nonlinear Pricing.* Oxford University Press.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 144–162

A Bridge from Monty Hall to the Hot Hand: The Principle of Restricted Choice

Joshua B. Miller and Adam Sanjurjo

Suppose that you work in a restaurant where two regular customers, Ann and Bob, are equally likely to come in for a meal. Further, you know that Ann is indifferent among the 10 items on the menu, whereas Bob strictly prefers the hamburger. While in the kitchen, you receive an order for a hamburger. Who is more likely to be the customer: Ann or Bob?

One intuition is that we have learned nothing from the observation that a hamburger was ordered, as it does not rule out either Ann or Bob, so they must remain equally likely to be the customer. However, this intuition is wrong, as it fails to account for *how* Ann and Bob choose items from the menu. By contrast, once we do account for how they choose, then the correct intuition emerges right away: because ordering a hamburger is more consistent with Bob (who must order it) than with Ann (who may order it), the order is more likely to have been placed by Bob.

While it may be easy to resist the incorrect intuition when confronting this simple problem, doing so is not so straightforward once the way that choices are made becomes even slightly less transparent. Let us briefly consider two examples: the Monty Hall problem and the presumed debunking of the "hot hand" phenomenon.

■ *Joshua B. Miller is Associate Professor of Economics, University of Melbourne, Melbourne, Australia. Adam Sanjurjo is Associate Professor of Economics, University of Alicante, Spain. Both authors contributed equally, with names listed in alphabetical order. Their email addresses are Joshua.Benjamin.Miller@gmail.com and sanjurjo@ua.es.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.144 doi=10.1257/jep.33.3.144

The Monty Hall problem is a probability puzzle known for its ability to confound the intuitions of both the layperson and the mathematically sophisticated. A standard version of the problem, taken from vos Savant (1990), is as follows:

> *Monty Hall problem: Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?*

While the intuitively appealing answer is that either of the two remaining doors leads to the same chances of winning the car, the chances actually increase if the contestant switches from door #1 to door #2 (under natural assumptions that we discuss later). People typically get this problem wrong. For example, a robust finding in laboratory experiments is that roughly 80–90 percent of subjects incorrectly stay with the same door, rather than switch (for example, Friedman 1998). Further, even a number of mathematically inclined academics (including Paul Erdős) have expressed disbelief when told the correct answer (Vazsonyi 1999).[1]

The hot hand fallacy refers to people's tendency to believe that success breeds success, even when it does not. In the seminal study by Gilovich, Vallone, and Tversky (1985), the authors found that basketball players shoot no better after having just made several shots in a row, despite a near-unanimous belief reported by players, coaches, and fans that players shoot better in these situations. When confronted with the scientific evidence against their beliefs, even professional players and coaches were left unpersuaded, leading the hot hand to become known as a "massive and widespread cognitive illusion" (Kahneman 2011).[2]

However, with the recent discovery of a surprising statistical bias (Miller and Sanjurjo 2018), it appears that the basketball community may have been right all along. In particular, to estimate a player's probability of making a shot, conditional on having made several in a row, Gilovich, Vallone, and Tversky (1985) and subsequent studies (1) selected the shot attempts that immediately followed a streak of several made shots (for example, three) and then (2) calculated the player's shooting percentage on these shots. As discussed below, this procedure biases the researcher toward overselecting missed shots, which leads to an underestimate of the player's probability of success on these shots. Not only is this *streak selection bias* large enough to invalidate the conclusions of previous studies, but it masks significant evidence of substantial hot hand shooting in their data.

[1] Math puzzles of this sort have been noted for their importance in stimulating research ideas and illustrating principles from microeconomic theory (Friedman 1998; Kluger and Wyatt 2004; Fehr and Tyran 2005). The Monty Hall problem, in particular, has been studied extensively, including in the first issue of this journal (Nalebuff 1987). For more discussion, see Rosenhouse (2009) and the references therein.

[2] The hot hand fallacy has been offered as a candidate explanation for certain puzzles and anomalies in financial markets, sports wagering, casino gambling, and lotteries. See Benjamin (2018), Miller and Sanjurjo (2018), Rabin and Vayanos (2010), and the references therein.

While it may not appear that there is any connection between why people have difficulty understanding the Monty Hall problem and why researchers had long overlooked the bias in common measures of the hot hand, we show that the two are in fact intimately related. The first step in understanding the relation is to observe that both environments involve a procedure that selects an observation for analysis on the basis of the outcomes of other observations in the same dataset. In particular, just as Monty offers the contestant an opportunity to switch to another door, knowing that a goat is behind the door he just opened, the hot hand researcher selects a shot from a longer sequence of basketball shots, knowing that the previous several shots were made. The key step to connecting these two environments, and many others, is then to illuminate the information that is revealed by their respective selection procedures.

The tool that we use to draw out these connections is the *principle of restricted choice*, an inferential rule drawn from the card game contract bridge that makes clear the information revealed by the optimizing behavior of a constrained opponent. The principle's simple intuition is illustrated above in the opening example with Ann and Bob, where Bob is more restricted to choose the hamburger than Ann is, because while Ann might order the hamburger, Bob must. In the next section, we show that restricted choice is naturally quantified as the updating factor from the odds formulation of Bayes' rule. To illustrate how intuitive and general restricted choice thinking is, we apply it to a number of settings. First, we use it to solve several classic probability paradoxes, including the Monty Hall problem.[3] This exercise makes clear that restricted choice renders intuitive the typically difficult counterfactual (and hypothetical) reasoning that is inherent in Bayesian updating. By contrast, we describe how some commonly used heuristic approaches, while helpful for particular problems, can lead to mistakes when applied more generally. We also use the principle to solve a progression of novel coin-flip probability puzzles, and to make comparisons across puzzles. For example, we show that one of our coin-flip puzzles captures the essence of the hot hand selection bias and at the same time is virtually equivalent to the Monty Hall problem.

Lastly, we consider various empirical examples in which restricted choice thinking can help researchers become aware of (and avoid) the types of counterintuitive mistakes and biases that can arise when particular observations are selected for analysis on the basis of the outcomes of other observations in the same dataset. Our four examples include (1) a bias that arises in measures of dependence across time, illustrated with the hot hand literature; (2) a bias that arises in measures of dependence across space, illustrated with Schelling's (1971) well-known work on segregation; (3) an unexpected correlation known as Berkson's paradox, illustrated with the canonical case of two unrelated diseases that happen to be negatively correlated in the hospitalized population despite being uncorrelated in the general

[3] Reese (1960, p. 29) illustrates the principle of restricted choice with a problem nearly identical to the Monty Hall problem. Gillman (1992) appears to be the first to use the restricted choice principle to explain the intuition behind the Monty Hall problem.

population; and (4) a hypothetical example of ESP research gone wrong. These examples are chosen to illustrate some pitfalls that researchers can avoid by using restricted choice thinking.

The Principle of Restricted Choice

The principle of restricted choice was first introduced in the context of the card game contract bridge, to account for the information revealed by the actions of an agent with a known decision rule. Legendary bridge player Terence Reese succinctly illustrates the principle in *Master Play in Contract Bridge* (Reese 1960, p. 26): "Since East could have played either card indifferently from K–Q, the fact that he has played one affords an indication that he does not hold the other."[4] Another illustration, which requires no familiarity with card games, is provided in our Ann and Bob example from the beginning of this paper. To reiterate, Bob is more restricted to choose the hamburger than Ann is, because while Ann may order the hamburger, Bob must. As a result, once we find out that the customer ordered a hamburger, we should shift our beliefs toward the customer being Bob rather than Ann.

The principle of restricted choice provides an informal *intuition* for why beliefs should shift in a particular direction upon the arrival of new information and calls to mind the essential qualitative feature of Bayesian updating. Namely, Bayes' rule requires that the odds in favor of a proposition increase upon the arrival of information that is more likely in the case that the proposition is true, or conversely, that the odds in favor of a proposition decrease upon the arrival of information that is less likely in the case that the proposition is true.

From here on, we represent uncertainty with odds rather than probabilities, as this simplifies the reasoning in the types of problems we discuss. For example, a proposition with a 3/5 probability of being true has 3/5 "chances" in its favor for every 2/5 chances against. Given this, the odds in favor of the proposition can be written as 3/5:2/5, or equivalently as 3/2:1 (by dividing each term by 2/5, as odds are invariant to proportional scaling). In turn, the odds of 3/2:1 can be stated simply as the single number 3/2, taking as given that the chances against the proposition are 1. Of course, associated probabilities can be easily recovered from the odds; for example, a proposition with 3:2 odds in its favor has 3 chances in its favor out of $3 + 2 = 5$ total chances—or a probability of 3/5.

To see how restricted choice can be understood as Bayesian updating, let A ("Ann") and B ("Bob") represent the two hypothetical propositions (or models) that could have produced the observed outcome c ("hamburger") in the restaurant

[4] Reese (1960, chap. 3, p. 26) credits Alan Truscott, who wrote the daily bridge column for the *New York Times* from 1964 to 2005, for introducing restricted choice to the bridge community in the 1950s. Prior to that, Borel and Chéron (1940) use the concept, at least implicitly, by applying Bayes' rule to calculate probabilities in bridge problems.

example. Then, given the prior odds, which consist of the chances in favor of B (relative to the chance in favor of A), Bayes' rule gives the posterior odds in favor of B (relative to one chance in favor of A):

$$\text{Posterior odds in favor of } B = \text{Likelihood ratio} \times \text{Prior odds in favor of } B.$$

The likelihood ratio, also known as the Bayes factor, represents the multiplicative factor by which the number of chances in favor of B increase, decrease, or stay the same upon observation of c.[5] For our purposes, it can be thought of as B's restrictedness relative to A's, that is, the degree to which B is more likely to produce outcome c than is A. The principle of restricted choice tells us that, upon observation of an outcome, the odds shift in the direction of the model that is more likely ("restricted") to produce that outcome.

To illustrate, in the Ann and Bob restaurant example, the prior odds in favor of Bob being the customer (relative to Ann) are 1:1. However, once a hamburger has been ordered, because Bob is more likely to order the hamburger than Ann is, the odds in favor of Bob must increase. In particular, if Ann is equally likely to order each of the 10 items, then because Bob orders the hamburger for sure, he is 10 times more restricted to choose the hamburger than Ann. Therefore, the odds in favor of the customer being Bob increase by a factor of 10 upon learning that the customer ordered a hamburger. Thus, the posterior odds in favor of Bob are 10:1. Finally, if we assume for simplicity that Ann and Bob are the only possible customers, then because there are 10 chances in favor of Bob for every 1 chance in favor of Ann, the probability that the hamburger order came from Bob is 10/11.

Restricted Choice in Some Classic Conditional Probability "Paradoxes"

We show how the simplicity and intuition of restricted choice reasoning extend to several related classic conditional probability puzzles that often tend to confound people's intuition. We start with Bertrand's box paradox (Bertrand 1889; Gorrochurn 2012; presentation below adapted from Rosenhouse 2009), then present two versions of the boy-or-girl paradox, and finally return to the Monty Hall problem.

> *Bertrand's box paradox: Three boxes are identical in external appearance. The first box contains two gold coins, the second two silver coins, and the third one gold coin and one silver coin. You choose a box at random and draw a coin. Suppose that you draw a gold coin. What is the probability that the other coin is also gold?*

[5] More formally, posterior odds satisfy ($\mathcal{R}_A^B(c) \times$ Prior chances in favor of B) : (Prior chances in favor of A), where $\mathcal{R}_A^B(c)$ is the likelihood ratio, or Bayes updating factor. The likelihood ratio is defined as the ratio of the probability of c conditional on B to its probability conditional on A, that is, $\mathcal{R}_A^B(c) = \frac{\Pr(c|B)}{\Pr(c|A)}$ (assuming $\Pr(c|A) > 0$). In the extreme case that $\Pr(c|A) = 0$, the odds in favor of B are 1:0 (assuming $\Pr(c|B) > 0$).

Given that a gold coin was drawn, it is impossible that the all-silver box was chosen. Thus, two possible boxes remain: all gold and mixed. Given this, it becomes intuitively appealing to conclude that the probability that the other coin is gold is 1/2. However, what this reasoning misses is that a draw that occurs from the all-gold box is more restricted to "choose" (draw) a gold coin, because with the gold box one *must* draw a gold coin, whereas with the mixed box one can draw either a gold or a silver coin. In this case, one is twice as restricted to choose the gold coin from the all-gold box relative to the mixed box. Therefore, by the principle of restricted choice, the updated odds in favor of the draw having come from the all-gold box are 2:1, double the prior odds of 1:1. This implies that the probability that the other coin is also gold is equal not to 1/2 but rather to 2/3.

Next, we consider the boy-or-girl paradox (as presented in problem 1 of Bar-Hillel and Falk 1982; see also Gardner 1961):

> *Boy-or-girl paradox: Mr. Smith is a father of two. We meet him walking along the street with a young boy whom he proudly introduces as his son. What is the probability that Mr. Smith's other child is also a boy?*

The intuitive answer to this problem is 1/2, and under usual assumptions, this answer is correct—but for reasons that differ from the intuition many people bring to the problem. Let us assume that Mr. Smith chooses his walking companion at random from among his two children (without discriminating). With this, the problem becomes close to Bertrand's box paradox: Mr. Smith's children are drawn, without replacement, from either an all-boy "box," an all-girl box, or a mixed-gender box. The key difference, however, is that the types of boxes are not all equally likely. In particular, the equivalent of the mixed box—one boy and one girl—has 2:1 prior odds in its favor, relative to any single-gender box, because there are two birth order possibilities in the mixed-gender box (boy–girl and girl–boy). Analogous to Bertrand's box paradox, learning that the randomly chosen walking companion is a boy makes the posterior odds in favor of both children being boys (relative to mixed gender) double the prior odds, because the choice of a boy is twice as restricted in the all-boys case. Thus, because the prior odds were 1:2 "in favor" of all boys (relative to mixed gender), or 1/2:1, the posterior odds are 1:1 in favor of all boys. Finally, because there remain only two possible compositions of children—all boys or mixed gender—the probability that Mr. Smith has all boys is 1/2. As a result, the probability that his other child is a boy is 1/2.[6]

[6] Another common version of the boy-or-girl paradox is as follows: "Mr. Smith says: 'I have two children and at least one of them is a boy.' Given this information, what is the probability that the other child is a boy?" (Fox and Levav 2004, p. 631). If one assumes that Mr. Smith would say nothing (or its equivalent) in the case that he were to have two girls, then in this version of the problem, Mr. Smith is equally restricted to report "boy" in the cases of boy–girl, girl–boy, and boy–boy, so prior and posterior odds are identical. As a result, the correct probabilities can be computed simply by an enumeration of the sample space and elimination of the impossible girl–girl combination. Therefore, failure to see the correct answer in this version of the boy-or-girl paradox can arise not because of a failure to incorporate

Now consider another version of the boy-or-girl paradox (equivalent to problem 2 in Bar-Hillel and Falk 1982, with slightly adapted language):

> *Younger boy-or-girl paradox: Mr. Smith is a father of two. We meet him walking along the street with a boy whom he proudly introduces as his eldest child. What is the probability that Mr. Smith's younger child is also a boy?*

Because the younger child must be either a boy or a girl, the intuitively appealing response to this question is, again, 1/2. This response is correct if we assume that Mr. Smith chooses his walking companion at random between his two children, regardless of gender, as in the basic boy-or-girl paradox.

But while gender neutrality is a natural assumption, another possibility is that Mr. Smith has the unfortunate attitude of being willing to walk only with sons. Assume that this is so, but that if he has two boys, then he is indifferent between walking companions and chooses one of the boys at random. Under these assumptions, observing the gender of Mr. Smith's walking companion yields redundant information. That is, if we had merely observed Mr. Smith walking with a child, without any further information, we would already know that the child must be a boy, and that the possible birth order combinations are thus boy–girl, girl–boy, and boy–boy.

However, in the current problem, we additionally discover that Mr. Smith's walking companion is his eldest child, which eliminates the possibility of boy–girl, reducing the possibilities to girl–boy and boy–boy. With this, the intuitive response is again 1/2, but now this response is wrong. The reason why is that it fails to take into account that the degree of restrictedness in Mr. Smith's choice varies across these hypothetical birth orders. In particular, if the younger child is a girl (girl–boy), then Mr. Smith's choice of walking partner will be the older boy for sure. On the other hand, if the younger child is also a boy (boy–boy), then Mr. Smith is equally likely to choose each boy. This means that when the younger child is a girl, Mr. Smith's choice is twice as restricted. Therefore, the posterior (relative) odds in favor of girl–boy are double the prior odds of 1:1. Thus, the probability of girl–boy is 2/3. As a result, the probability of boy–boy is 1/3—that is, there is a 1/3 chance that the younger child is a boy.

We now return to the Monty Hall problem, which is essentially identical to the younger boy-or-girl paradox just discussed, in which Mr. Smith is willing to walk only with sons. With respect to the statement of the problem in vos Savant (1990; see also Selvin 1975), we change the door numbers (without loss of generality) in order to facilitate comparison with the coin-flip problems presented below:

the subtleties of Bayesian reasoning, but simply because of a failure to appreciate the subtleties of the sample space. A classic example of this type of mistake is Leibniz's error, which is believing that 11 and 12 are equally probable when rolling a pair of fair dice, because there is just one way for each sum to be partitioned into two numbers less than (or equal to) 6 (Gorroochurn 2012).

> *Monty Hall problem: Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #3, and the host, who knows what's behind the doors, opens another door, say #1, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?*

As with the boy-or-girl paradox, the correct answer depends on conditions that have not yet been specified. One possibility is that Monty, the host, follows a rule that he must always reveal a goat from behind one of the two doors that the contestant does not choose. Further, in the case that Monty has two goats to choose from, he chooses a door (uniformly) at random.

Under these conditions, because one goat and one car will always remain covered once Monty reveals a goat, an intuitively appealing conclusion is that the odds in favor of the car being behind door #2 are 1:1 (relative to door #3), meaning that the contestant should be indifferent about switching.

Nevertheless, as in the previous problems, this simple reasoning is incorrect. To see why, notice first that before Monty opens door #1, the contents behind doors #1 and #2, respectively, are one of the following, each with equal probability: car–goat, goat–car, or goat–goat. However, once Monty reveals a goat behind door #1, the remaining possible arrangements behind doors #1 and #2 become goat–car and goat–goat. Because Monty must open door #1 in the case of goat–car, whereas he opens it only half of the time in the case of goat–goat, he is twice as restricted to open it in the case of goat–car. Therefore, given that the prior (relative) odds in favor of goat–car were 1:1, the posterior odds must double—that is, the odds in favor of the car being behind door #2 are now 2:1 (relative to door #3). As a result, it is in the contestant's interests to switch doors, as the probability of winning the car by doing so is 2/3.

Restricted Choice as a General-Purpose Approach

Throughout this paper we illustrate how the restricted choice approach is intuitive and straightforward to apply to a range of conditional probability problems. By contrast, while other approaches can do an excellent job of shaking people out of incorrect initial intuitions, they tend to employ either ad hoc explanations that do not readily generalize across problems or formal explanations that do, but at the expense of being less intuitive.

For example, in the *Parade Magazine* article in which she discussed the Monty Hall problem, vos Savant (1990) offered a modification of the problem to make more salient the benefit of switching after Monty opens a door to reveal a goat. She wrote, "Here's a good way to visualize what happened. Suppose there are a million doors, and you pick door #1. Then the host, who knows what's behind the doors and will always avoid the one with the prize, opens them all except door #777,777. You'd switch to that door pretty fast, wouldn't you?"

This modification effectively conveys the restricted choice intuition in a way that helps make the correct answer—to switch doors—more transparent. In the terms we have been using, because Monty must leave door #777,777 closed when the car is behind it, whereas he has a 1/999,999 probability of leaving it closed when the car is behind door #1, he is 999,999 times more restricted to leave door #777,777 closed when the car is behind door #777,777 (versus door #1). Because the prior odds between the two doors are 1:1, the posterior odds become 999,999:1 in favor of door #777,777. Indeed, when experimental subjects face a many-door version of the Monty Hall problem, they correctly decide to switch doors at a rate of approximately 85 percent, compared with only 15 percent when facing the standard version (Page 1998).

While the many-doors modification of the Monty Hall problem does lead to an immediate improvement in the rate of correct responses, it also has some important limitations. For one, when experimental subjects who face the manipulation then go back to the standard version of the Monty Hall problem, they proceed to make the wrong choice at rates similar to subjects that never faced the many-door version (Page 1998). Second, it does not indicate how to compute the posterior odds, which is necessary if one wishes to ascertain the value of switching. Third, the modification seems unlikely to be useful as a general problem-solving tool, as it is difficult to adapt to the other problems we have discussed so far.[7]

Another common approach to solving the Monty Hall problem—and the highest-voted answer on the question-and-answer website Mathematics Stack Exchange (https://math.stackexchange.com/q/96832)—involves answering as if the contestant decides whether to commit to switching *before* Monty chooses which of the two remaining doors to open (see also Krauss and Wang 2003). While this heuristic approach answers a slightly different problem, it appears to help people see that always switching yields the best of what the two remaining doors have to offer, and thus yields the car 2/3 of the time.

While reasoning through the Monty Hall problem without conditioning on which door Monty opens may help people shake off certain incorrect intuitions, this best-of-two-doors approach also has some important limitations. For one, it is not clear how to generalize it to address the other conditional probability problems discussed in the previous section. More importantly, because the best-of-two-doors approach ignores which of the two doors was opened, as well as Monty's rule for choosing between them in the case of two goats, the resulting probability—while correct numerically—is not the *conditional* probability that the problem implicitly requests. To see why this matters, assume that in the case that Monty has two goats to choose between, he always reveals the goat behind the lower-numbered door (rather than randomizing between the two doors, as implicitly assumed above). While the best-of-two-doors intuition still indicates that it is always strictly beneficial

[7] In the boy-or-girl problems, the analogous modification is for Mr. Smith to walk with all but one of his 999,999 children, and to meet him walking with only boys. In Bertrand's box paradox there would be 999,999 coins in each box, 999,998 coins would be drawn, and they would all need to be gold.

for the contestant to switch, this is no longer true in the event that Monty opens the lower-numbered door! Instead, the contestant should be indifferent between switching and not switching because Monty is now equally restricted to open the lower-numbered door, regardless of whether his two options are goat–goat or goat–car. Finally, while one could claim that this argument is unnatural, because Monty should be expected to randomize uniformly in the case of two goats, in the next section we provide an example of a coin-flip version of the Monty Hall problem in which the best-of-two-doors intuition fails to provide the correct answer even when Monty does randomize uniformly.

Yet another approach to solving conditional probability problems is to describe the sample space in detail and calculate the conditional probability directly. In the Monty Hall problem, for example, given the contestant's initial choice, one can generate all four (prize-placement, door-opened) combinations, and their probabilities, by laying out Monty's two-stage decision tree in which he first places the car behind one of the three doors (at random) and then chooses which door to open (according to his rule). One can then grind out the correct answer using the definition of conditional probability, rather than Bayes' rule.[8] While certainly correct, the relative disadvantage of sample space arguments is that they are typically more complex, and the intuition is less transparent.

When it comes to conditional probability problems, ad hoc intuitive explanations—as well as more complicated formal explanations—may be correct as far as they go. However, they are limited relative to restricted choice in terms of building a broader intuition for how the probability of interest in these kinds of problems can be altered by seemingly small changes in the selection procedure.

Restricted Choice in Coin-Flip Puzzles

In this section, we introduce a progression of coin-flip puzzles ("paradoxes") and solve them using restricted choice reasoning. The *next flip paradox* is nearly identical to the Monty Hall problem. When combined with the *alternation paradox*, it provides an explanation of why the earlier studies that purported to demonstrate a hot hand fallacy were actually biased. We then extend the alternation paradox into the *streak-reversal paradox*, which illustrates how these statistical puzzles can be related to selection bias in slightly richer settings.

> *Next flip paradox: Jack flips a coin three times, then tells you that the first flip is a heads. What is the probability that the second flip is also a heads?*

[8] More broadly, one can use a natural frequency intuition to arrive at the correct conditional probabilities for the Monty Hall problem. Gigerenzer and Hoffrage (1995) adapt the natural sampling approach to reframe conditional probability problems so that subjects can apply the definition directly, rather than updating priors with Bayes' rule.

The answer to this question depends on conditions that have not yet been specified. In particular, if Jack had decided to select which of the first two flip outcomes to reveal at random, or had simply planned on always revealing the outcome of the first flip, then the correct answer will be 1/2. This is precisely as in the basic boy-or-girl paradox, in which Mr. Smith chooses a child at random, regardless of gender. But instead, say that Jack was interested only in the respondent's beliefs about the probability that heads follows heads. Thus, assume that Jack had selected one of the first two flip outcomes at random according to the criterion that it be a heads (so that with two tails he could not have asked the question). In this case, the answer changes.

Under this selection criterion, the next flip paradox is nearly identical to the standard Monty Hall problem. In particular, just as Monty is able to look behind each door before opening one, which in turn reveals information regarding the location of the car, Jack looks at the outcome of each coin flip before selecting one, which in turn reveals information about the location of heads. To see the parallel more clearly, let Jack now be the game show host instead of Monty. In this game, Jacks flips three coins, leaving each behind a separate door. He then asks the contestant to choose one of the three doors, informing her that she will receive a prize if the door she chooses conceals a tails flip. Once the contestant has chosen one of the doors, Jack opens one of the other two doors at random, according to the criterion that he must reveal a heads flip (in the case of two tails flips, he cannot open either door). Finally, Jack offers the contestant the opportunity to switch. Assume that the contestant's initial choice is the third door, and that Jack opens the first door, revealing the first flip to be a heads. In this case, the first two flip outcomes must be either heads–tails or heads–heads. Then, by the same restricted choice reasoning as in the Monty Hall problem, Jack is twice as restricted to open the first door in the case of heads–tails as he is in the case of heads–heads. As a result, the probability that the second flip is a heads is 1/3.[9]

Although the contestant can extract information about the second coin flip from the knowledge that the first flip is a heads, this does not imply that coins have memory. Instead, the contestant exploits the fact that Jack has inspected the outcome of the first two flips before choosing, which means that Jack's choice (probabilistically) reflects his knowledge. More subtly, this also implies that time's arrow is irrelevant—that is, if Jack were to instead reveal that the second flip was a heads, then the probability of heads on the previous (first) flip would similarly be 1/3.

Another coin-flip problem, the alternation paradox, brings us one step closer to illustrating the streak selection bias; indeed, this problem happens to be the exact probabilistic representation of the simple three-flip example of the bias given in table 1 of Miller and Sanjurjo (2018).

[9] A slight modification makes the next flip paradox identical to the younger boy-or-girl paradox: in this version, Jack flips the coin twice, then chooses one of the heads flips at random (final flip included) and tells you that it is the first flip.

> *Alternation paradox: Jack will flip a coin three times, then select a flip that is immediately preceded by a heads, at random. Assuming that Jack has a flip to select, what is the probability that the selected flip is a heads?*

In order for Jack to select a flip, he must inspect the outcomes of the first two flips. Given that at least one of the two has come up heads, it is clearly impossible that the sequence could have started with two tails. Let H__ be the event that Jack selects the second flip, which is preceded by a heads on the first flip; let _H_ be the event that Jack selects the third flip, which is preceded by a heads on the second flip. Conditional on Jack having chosen a flip, these events are equally likely. In the case that Jack selects the second flip, the probability that it is a heads is simply the solution to the next flip paradox, namely, Pr(HH_|H__) = 1/3. On the other hand, if Jack selects the third flip, then Pr(_HH|_H_) = Pr(__H) = 1/2, as the outcome of the last flip cannot restrict Jack's choice of which immediate heads successor to select. It then immediately follows that

$$
\begin{aligned}
&\text{Pr(Heads|Flip preceded by a heads)}\\
&\quad = \text{Pr(HH\_|H\_\_)} \times \text{Pr(H\_\_)} + \text{Pr(\_HH|\_H\_)} \times \text{Pr(\_H\_)}\\
&\quad = \left(\frac{1}{3} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right)\\
&\quad = \frac{5}{12}.
\end{aligned}
$$

The next problem extends the alternation paradox to 100 flips and streak lengths of 3.

> *Streak reversal paradox: Jack, now a researcher, observes the outcome of 100 flips of a fair coin. He selects all of the flips that are immediately preceded by three consecutive heads and calculates the proportion of heads on these flips. He expects this proportion to be 0.5. Is he correct?*

While Jack's expectation is intuitively appealing, it turns out to be incorrect. In particular, the expected value of this proportion is not 0.50 but 0.46 (for the formula, see Miller and Sanjurjo 2018).

To see how the principle of restricted choice provides intuition for the streak reversal paradox, first observe that the expected proportion can be represented as a probability. In particular, the proportion of heads among the flips that Jack has selected is equal to the probability of heads on a flip chosen at random from among these flips. Next, imagine Jack choosing a flip at random from among the flips that he selected (those immediately preceded by three consecutive heads). If Jack were to choose, say, flip number 42, then intuition suggests that the odds of heads on that flip are 1:1. As with the alternation paradox, this intuition would be correct if Jack were to have chosen the flip *before* having examined the sequence. However, because

Jack instead examined the sequence first, then chose flip 42 based on information that he had regarding the outcomes of other flips in the sequence (including flip 42), this intuition is incorrect.

To see why the odds of heads on flip 42 are not 1:1, first observe that if flip 42 were a heads, then flips 39–42 would be HHHH, making flip 43 also immediately follow (at least) three consecutive heads. In this case Jack could have chosen flip 43 instead of flip 42. On the other hand, if flip 42 were instead a tails, then flips 39–42 would be HHHT, making it impossible for Jack to choose flip 43 (or 44, or 45). This implies that with a tails on flip 42 Jack would be relatively more restricted (likely) to choose flip 42, as there would be comparatively fewer eligible flips (on average) in the sequence from which to choose. Finally, the fact that Jack is more restricted to choose flip 42 in the case that it is a tails makes the likelihood that the flip he chose was a tails greater than the unconditional (prior) probability of flipping a tails, which in turn implies that the (posterior) probability that flip 42 is a heads is less than 0.5.[10]

This reasoning holds for any flip that Jack may choose, unless it happens to be the final flip of the sequence. For that flip, the posterior odds of a heads versus a tails are the same as the prior odds for the same reason given in the explanation of the alternation paradox.

Some Empirical Implications

We provide four empirical examples of how applying the principle of restricted choice can in some instances help us as researchers to avoid making critical mistakes in our design of experiments, analysis of data, and interpretation of results.

The Presumed Debunking of the Hot Hand

Having gone through the solutions to the coin-flip puzzles, it is now straightforward to explain the bias built into the seminal study of the hot hand fallacy by Gilovich, Vallone, and Tversky (1985) and similar studies that followed.

The original study conducted a controlled shooting experiment in which collegiate basketball players attempted 100 shots, from locations on the court at which they are expected to make half of them. To test for a hot hand, the authors compared each player's shooting percentage immediately following a streak of successes (makes) with his/her percentage immediately following a streak of failures (misses). Under their null hypothesis of no hot hand shooting, these two percentages are expected to be the same, and under the alternative hypothesis of hot hand shooting, the percentage following successes is expected to be larger than the percentage following failures.

[10] This explanation omits some details; see appendix A of Miller and Sanjurjo (2018) for a complete proof.

While this null hypothesis may seem correct, the streak reversal paradox makes clear that, perhaps counterintuitively, it is not. Indeed, if a robot player's shot outcomes were to be determined by repeated tosses of a fair coin (no hot hand), the expected shooting percentage following streaks of success would not be 0.50, but 0.46. By symmetry, the expected percentage following streaks of failures would be 0.54. Taking the difference, the total bias is 8 percentage points.[11] This means that if a researcher were to observe no difference in a player's shooting percentages, it would actually constitute (sizeable) evidence of the hot hand!

Upon correction for this bias in the shooting percentages of each of the players in the original study, the positive 3 percentage point average hot hand effect reported there (not statistically significant) becomes a statistically significant 13 percentage point effect (Miller and Sanjurjo 2018). This is a large effect and is roughly equal to the difference between a median and a top three-point shooter in the 2015–2016 NBA season.

Similarly biased measures were also used in the replications of the original hot hand study: a close replication with Olympic basketball players (Avugos, Bar-Eli, Ritov, and Sher 2013) and another using elite shooters from the annual NBA three-point "shootout" (Koehler and Conley 2003). As with the original study, a bias-corrected reanalysis reveals substantial evidence of hot hand shooting in both datasets (Miller and Sanjurjo 2018).

Clustering and Segregation

While the coin-flip puzzles and the streak selection bias pertain to measures of sequential dependence in time-series data, it turns out that the time dimension itself is not central to the bias. Instead, the key is that the selection of the data to be analyzed is determined by the outcomes of other (adjacent) flips in the same dataset. This in turn suggests the possibility of a more general selection bias that applies to measures of dependence across space just as easily as is does to measures of dependence across time.

Consider an $n \times n$ grid of cells, each colored red or blue according to the outcome of a fair coin flip. Now, suppose that we are interested in the probability that a cell is a red, given that all of its neighbors are red. An intuitive way to estimate this probability would be to select the subset of all cells that are surrounded by red and then calculate the proportion of red among these cells. However, this estimate will be biased downward due to a mechanism that is essentially identical to the bias that emerges in the one-dimensional setting of the streak reversal paradox. In particular, if one were to choose a cell from among those surrounded by red, the probability that this cell is blue would be greater than 50 percent. This is because if it were blue, then none of its neighbors could be surrounded by red, which would

[11] In fact, the bias is actually a bit more severe than this, due to an additional selection effect that is driven by the exclusion of sequences that do not have both of the following: (1) at least one shot that immediately follows a streak of made shots and (2) at least one shot that immediately follows a streak of missed shots.

lead to fewer such cells, making the probability of choosing any such cell, including itself, more likely.

The bias in this measure of the similarity between a cell and its neighbors suggests the possibility of such a bias appearing in measures of clustering in location preferences, as in studies of racial segregation. Indeed, this description of a grid of cells with two possible values is reminiscent of the classic work by Schelling (1971) on patterns of segregation. While it has no bearing on Schelling's main results, a bias happens to exist in one of his measures of clustering—"the average proportion of neighbors of like or opposite color." The reason for the bias is similar to that described in the previous paragraph. In particular, imagine choosing a cell at random from among the red cells. If more of that cell's neighbors are blue, then fewer red cells are available to be drawn. This in turn makes the chosen cell more likely to have been chosen to begin with. By consequence, a representative red cell is expected to have a higher proportion of blue neighbors than red neighbors.[12]

This bias extends to any spatial arrangement of outcomes, including lattices and networks. It is closely related to a bias in a well-known measure of spatial association, Moran's I (Moran 1950). The extent to which the magnitude of these biases is empirically relevant depends on the definition of a cluster and the size of the grid under consideration.[13]

Berkson's Paradox

Berkson's paradox (sometimes called Berkson's bias) is a form of selection bias. The original example involved a hypothetical case of two diseases that, while not associated in the general population, become negatively associated in the population of hospitalized patients (Berkson 1946). It is sometimes referred to as the "admission rate bias" (Sackett 1979), or as an instance of "collider bias" (for example, Westreich 2012), and it can be illustrated with the following example (adapted from Pearl 2009):

> *Berkson's paradox: Suppose that a randomly selected high school student has a 50 percent chance of having good SAT scores, along with a 50 percent chance of having good grades, and that the attributes are independent. Further, suppose that every student*

[12] Schelling (1971, p. 156) briefly considers this biased measure of segregation. Specifically, Schelling writes, "If we count neighbors of like color and opposite color for each of the 138 randomly distributed stars and zeros in [Schelling's figure 7], we find that zeros on the average have 53 percent of their neighbors of the same color, stars 46 percent. (The percentages can differ because stars and zeros can have different numbers of blank neighboring spaces.)" Of course, Schelling's main result was not to measure segregation but rather to show that a relatively weak preference for being near one's own type, together with the possibility of movement, would often lead to much stronger patterns of segregation.

[13] The bias in Moran's I measure of spatial autocorrelation is typically small, with an expected value of $-1/(n-1)$, where n is the total number of cells. For the cluster-related measures of association discussed above, the bias is stronger, but still weaker than the streak-related measures in time-series data. For example, in a 50 × 50 grid, the probability that one of the cells surrounded by 8 reds is itself red is approximately 48 percent, whereas in a 2,500-cell linear grid, the probability that one of the cells with 8 consecutive red cells to its left is itself red is approximately 44 percent.

with at least one good attribute applies to university and highlights his/her single best attribute in the application. If an applicant highlights good grades, then what is the probability that the applicant has good SAT scores?

Assuming that an applicant highlights an attribute at random (uniformly) in the case that both attributes are good, this problem is identical to the younger boy-or-girl paradox, as well as a two-coin version of the next flip paradox. In this problem, each attribute is good or not good with a 50–50 chance, just as each child is a boy or not a boy with a 50–50 chance. As a result, an applicant who has good grades and poor SAT scores is twice as restricted to highlight good grades compared with an applicant with both good grades and good SAT scores. Thus, as prior odds are even, the principle of restricted choice leads to posterior odds of 2:1 in favor of the applicant having good grades and poor SAT scores; that is, given an emphasis on good grades, the probability that the applicant has poor SAT scores is 2/3.

While it remains true that, among the applicants with good grades, half of them also have good SAT scores, this subgroup constitutes just 1/3 of the applicant pool. The remaining 2/3 of the applicants, on the other hand, have just one good attribute. Thus, there will be a negative correlation between attributes in the applicant pool, despite the correlation in the general population being zero.

This phenomenon could easily lead a casual observer to fallacious beliefs. For example, a student (or professor) who spends enough time in a university environment may come to believe (incorrectly) that certain attributes that are associated with good grades (like diligence) are in general inversely related to those attributes associated with good SAT scores (like brilliance). This mistake is analogous to a gambler holding the belief that streaks are more likely to end rather than continue, because in his personal experience this is, in fact, representative of a typical night at the casino (as conveyed in the streak reversal paradox, presented above, and in gambler's verity, presented below).

It is not difficult to imagine that a similar bias may be present in experiments in which performance on behavioral tasks that involve cognitive ability is correlated with a personality measure such as conscientiousness. Indeed, because experimental subjects in research studies may be further selected on attributes such as budget constraints and intellectual curiosity, one can similarly imagine the discovery of appealing new correlations that are nevertheless spurious—such as a hypothetical negative correlation between measures of intellectual curiosity and greedy or selfish behavior in experimental tasks. As one example, Murray, Johnson, McGue, and Iacono (2014) proposed that empirical work documenting an (internally valid) negative correlation between conscientiousness and cognitive ability may instead merely be reporting a statistical artifact that is driven by a selection bias identical to Berkson's paradox.

A Hypothetical Case: Gambler's Verity and Psi Research

The same bias that underlies the alternation paradox can be used to generate a puzzle in which a strategy for predicting randomly generated outcomes can appear to outperform what would be expected by chance. In particular, this can happen if a researcher is unaware of the implicit selection bias that the strategy generates.

> *Gambler's verity: Imagine a roulette wheel in which half of the slots are red and half are black (for simplicity). Jill will observe exactly three spins of the wheel and has committed to the following betting strategy: whenever observing a red (R), bet black (B) on the next spin; otherwise, do not bet. Do you expect Jill to win half of her bets?*

Jill's betting strategy will restrict her to betting on the second spin, the third spin, or both. Thus, there are three possible outcomes: she will win on none of her bets, half of them, or all of them. While intuition may suggest that she is expected to win on half of her bets, this is incorrect, as it overlooks the fact that the three outcomes are not equally likely. To see this, we can enumerate the sample space, as follows: if the sequence is BBB or BBR, Jill will not bet; otherwise, for the remaining six equally likely sequences, she will bet. Given that Jill bets, she has a 3/6 = 1/2 probability of winning all of her bets (RBR, RBB, BRB), a 1/6 probability of winning half of them (RRB), and a 2/6 = 1/3 probability of losing all of them (BRR, RRR). As a result, Jill is expected to win more bets than she loses, with an expected win rate of (1/2 × 1) + (1/6 × 1/2) + (1/3 × 0) = 0.58. In fact, her high success rate immediately follows from the solution to the alternation paradox, which we solved using the same restricted choice thinking as in our solution to the Monty Hall problem. That is, Jill's expected win rate is equivalent to the statement that for a randomly selected flip that is immediately preceded by a heads, the probability of a tails (an alternation) is 1 – 5/12 = 0.58.

While it appears that Jill has discovered a strategy with which she can expect to win money, this is not true. In particular, relative to the high-probability sequences in which she walks away ahead, in the low-probability sequences in which she walks away behind she wagers 50 percent more and her absolute (negative) profit is 50 percent greater. The key to this asymmetry is that in some of these sequences Jill is betting only once, but in others she is betting twice. Specifically, conditional on walking away ahead, the sequences RBR, RBB, and BRB are equally likely, and in each sequence Jill wagers once and wins once. On the other hand, conditional on walking away behind, while the sequences BRR and RRR are also equally likely, for the sequence BRR Jill wagers once and loses, but for the sequence RRR she wagers twice and loses twice. As a result, when Jill walks away behind (1/3 probability), she is expected to wager 1.5 times with a net payoff of –1.5, whereas when she walks away ahead (1/2 probability), she is expected to wager 1 time with a net payoff of 1. As a result, given fair odds, Jill is expected to break even. This (sad) state of affairs brings to mind the old Las Vegas proverb: the probability of winning is inversely proportional to the amount of the wager.

To see how the gambler's verity problem could have implications for social science research, consider the hypothetical case of the amazing Zener, an ESP master who claims to have a scientifically validated method to train people in precognition. In order to validate his method, he devises a test to prove that his students can do better than chance at predicting the outcomes of coin flips. For each student, a group of objective third-party researchers will flip a coin 100 times, and the student will predict only on flips for which he/she "senses" the ensuing outcome. According to Zener, not all of his trainees have learned how to predict, so he requests that the researchers merely count how many of his students predict at better than chance rates.

Following these instructions, the researchers find that of the 1,000 students tested, 490 predict at a rate better than chance, 395 at a rate worse than chance, and 115 at the rate of chance. Thus, the odds are found to be substantially in favor of a student predicting at better-than-chance rates, relative to worse-than-chance rates. Furthermore, the average student is observed to have a 54 percent success rate on his/her predictions. Mystified by these statistically significant results, the researchers are left to conclude that Zener must indeed have amazing abilities.

However, the researchers' conclusion is premature, as the observed results can easily occur in the absence of precognition. In fact, this outcome is close to what would be *expected* if Zener had instructed his students to simply predict a tails whenever the previous three flips are heads—the equivalent of predicting a streak reversal (tails) in the setting of the streak reversal paradox.

Conclusion

We have shown that the usefulness of the principle of restricted choice as an inferential tool extends well beyond the settings of contract bridge and the Monty Hall problem. When naturally quantified as the updating factor in the odds form of Bayes' rule, restricted choice provides a simple, intuitive, and general approach to thinking through and solving classic conditional probability puzzles. Moreover, it can be used to identify novel biases in important empirical settings. Thus, the principle is capable of helping researchers avoid certain intuitively appealing but critical errors when designing experiments, analyzing data, and interpreting results.

■ *Financial support from Bocconi University, Bocconi Experimental Laboratory for the Social Sciences (BELSS), and the Spanish Ministry of Economics and Competitiveness (Project ECO2015-65820-P) is gratefully acknowledged. We also gratefully acknowledge Generalitat Valenciana (Research Projects Grupos 3/086 and PROMETEO/2013/037). This draft has benefitted from conversations with Maya Bar-Hillel, Ben Cohen, Lola Collado, Vincent Crawford, Carlos Cueva, Martin Dufwenberg, Jordan Ellenberg, Maria Glymour, Iñigo Iturbe, Steve Kass, Mike Lipkin, Massimo Marinacci, Judea Pearl, Lones Smith, and Dan Stone, as well as seminar participants at University of Alicante, Radboud University Nijmegen, Nova School of Business and Economics, Technical University of Dortmund, the 2016 International Meeting on Experimental and Behavioral Social Sciences (IMEBESS), and the 2017 Research Unit on Complexity and Economics (UECE) Lisbon Meetings in Game Theory and Applications.*

References

Avugos, Simcha, Michael Bar-Eli, Ilana Ritov, and Eran Sher. 2013. "The Elusive Reality of Efficacy–Performance Cycles in Basketball Shooting: Analysis of Players' Performance under Invariant Conditions." *International Journal of Sport and Exercise Psychology* 11(2): 184–202.

Bar-Hillel, Maya, and Ruma Falk. 1982. "Some Teasers concerning Conditional Probability." *Cognition* 11(2): 109–22.

Benjamin, Daniel J. 2019. "Errors in Probabilistic Reasoning and Judgment Biases." Chap. 2 in *Handbook of Behavioral Economics*, vol. 2, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson. Amsterdam: Elsevier.

Berkson, Joseph. 1946. "Limitations of the Application of Fourfold Table Analysis to Hospital Data." *Biometrics Bulletin* 2(3): 47–53.

Bertrand, Joseph. 1889. *Calcul des probabilités*. Paris: Gautier-Villars et Fils.

Borel, Émile, and André Chéron. 1940. *Théorie mathématique du bridge à la portée de tous*. Paris: Gauthier-Villars.

Fehr, Ernst, and Jean-Robert Tyran. 2005. "Individual Irrationality and Aggregate Outcomes." *Journal of Economic Perspectives* 19(4): 43–66.

Fox, Craig R., and Jonathan Levav. 2004. "Partition-Edit-Count: Naive Extensional Reasoning in Judgment of Conditional Probability." *Journal of Experimental Psychology: General* 133(4): 626–42.

Friedman, Daniel. 1998. "Monty Hall's Three Doors: Construction and Deconstruction of a Choice Anomaly." *American Economic Review* 88(4): 933–46.

Gardner, Martin. 1961. *The Second Scientific American Book of Mathematical Puzzles and Diversions*. New York: Simon and Schuster.

Gigerenzer, Gerd, and Ulrich Hoffrage. 1995. "How to Improve Bayesian Reasoning without Instruction: Frequency Formats." *Psychological Review* 102(4): 684–704.

Gillman, Leonard. 1992. "The Car and the Goats." *American Mathematical Monthly* 99(1): 3–7.

Gilovich, Thomas, Robert Vallone, and Amos Tversky. 1985. "The Hot Hand in Basketball: On the Misperception of Random Sequences." *Cognitive Psychology* 17(3): 295–314.

Gorroochurn, Prakash. 2012. *Classic Problems of Probability*. New Jersey: John Wiley and Sons.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kluger, Brian D., and Steve B. Wyatt. 2004. "Are Judgment Errors Reflected in Market Prices and Allocations? Experimental Evidence Based on the Monty Hall Problem." *Journal of Finance* 59(3): 969–98.

Koehler, Jonathan J., and Caryn A. Conley. 2003. "The 'Hot Hand' Myth in Professional Basketball." *Journal of Sport and Exercise Psychology* 25(2): 253–59.

Krauss, Stefan, and X. T. Wang. 2003. "The Psychology of the Monty Hall Problem: Discovering Psychological Mechanisms for Solving a Tenacious Brain Teaser." *Journal of Experimental Psychology: General* 132(1): 3–22.

Miller, Joshua B., and Adam Sanjurjo. 2018. "Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers." *Econometrica* 86(6): 2019–47.

Moran, P. A. P. 1950. "Notes on Continuous Stochastic Phenomena." *Biometrika* 37(1/2): 17–23.

Murray, Aja L., Wendy Johnson, Matt McGue, and William G. Iacono. 2014. "How Are Conscientiousness and Cognitive Ability Related to One Another? A Re-Examination of the Intelligence Compensation Hypothesis." *Personality and Individual Differences* 70: 17–22.

Nalebuff, Barry. 1987. "Puzzles: Noisy Prisoners, Manhattan Locations, and More." *Journal of Economic Perspectives* 1(1): 185–91.

Page, Scott E. 1998. "Let's Make a Deal." *Economics Letters* 61(2): 175–80.

Pearl, Judea. 2009. "Causal Inference in Statistics: An Overview." *Statistical Surveys* 3: 96–146.

Rabin, Matthew, and Dimitri Vayanos. 2010. "The Gambler's and Hot-Hand Fallacies: Theory and Applications." *Review of Economic Studies* 77(2): 730–78.

Reese, Terence. 1960. *Master Play in Contract Bridge*. New York: Dover Publications.

Rosenhouse, Jason. 2009. *The Monty Hall Problem: The Remarkable Story of Math's Most Contentious Brain Teaser*. New York: Oxford University Press.

Sackett, David L. 1979. "Bias in Analytic Research." *Journal of Chronic Diseases* 32(1/2): 51–63.

Schelling, Thomas C. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology* 1(2): 143–86.

Selvin, Steve. 1975. "Letters to the Editor: A Problem in Probability." *American Statistician* 29(1): 67.

Vazsonyi, Andrew. 1999. "Which Door Has the Cadillac?" *Decision Line* 30(1): 17–19.

vos Savant, Marilyn. 1990. "Ask Marilyn." *Parade Magazine*, September 1990, p. 15.

Westreich, Daniel. 2012. "Berkson's Bias, Selection Bias, and Missing Data." *Epidemiology* 23(1): 159.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 163–184

A Toolkit of Policies to Promote Innovation

Nicholas Bloom, John Van Reenen,
and Heidi Williams

The US economy has experienced a slowdown in productivity growth since the 1970s, which—except for an upward blip between 1996 and 2004—has been remarkably persistent. Other developed countries have also experienced this disappointing productivity trend. Moreover, slow productivity growth has been accompanied by disappointing real wage growth for most US workers, as well as rising wage inequality.

Innovation is the only way for the most developed countries to secure sustainable long-run productivity growth. For nations farther from the technological frontier, catch-up growth is a viable option, but this cannot be the case for leading-edge economies such as the United States, Japan, and the nations of Western Europe. For countries such as these, what are the most effective policies for stimulating technological innovation?

In this article, we take a practical approach to addressing this question. If a policymaker came to us with a fixed budget of financial and political capital to invest in innovation policy, what would we advise? We discuss a number of the main innovation policy levers and describe the available evidence on their effectiveness: tax policies to favor research and development, government research grants, policies aimed at increasing the supply of human capital focused on innovation, intellectual

■ *Nicholas Bloom is William D. Eberle Professor of Economics, Stanford University, Stanford, California. John Van Reenen is Gordon Y. Billard Professor in Management and Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Heidi Williams is Professor of Economics, Stanford University, Stanford, California. All three authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are nbloom@stanford.edu, vanreene@mit.edu, and hlwill@stanford.edu.*

[†] For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.163

doi=10.1257/jep.33.3.163

property policies, and pro-competitive policies. In the conclusion, we synthesize this evidence into a single-page "toolkit," in which we rank policies in terms of the quality and implications of the available evidence and the policies' overall impact from a social cost-benefit perspective. We also score policies in terms of their speed and likely distributional effects.

We do not claim that innovation policy is the *only* solution to America's productivity problem. Indeed, even within the United States, many firms are well behind the technological frontier, and helping these firms catch up—for example, by improving management practices—would likely have very high value. Nonetheless, we believe that sensible innovation policy design is a key part of the solution for revitalizing leading economies and will lead to large long-run increases in welfare. Before beginning our tour, we start with some background facts and then address an obvious question: why should a policymaker spend any resources at all on innovation?

Some Background Facts

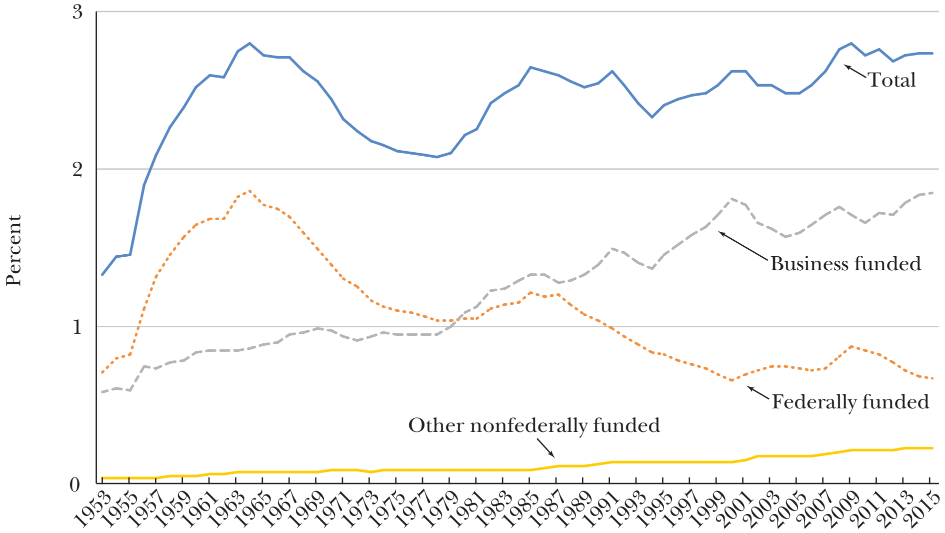
In 2015, spending on research and development (R&D) performed in the United States stood at just over $495 billion.[1] Figure 1 shows how this amount has evolved over time since 1953, in total as well as separately for R&D funded by businesses, the federal government, and other institutions (including state and local governments), as a share of GDP. R&D spending as a share of GDP grew from around 1.3 percent in 1953 to around 2.7 percent in 2015. Over time, there has been a relative decline in the share of R&D funded by the federal government, and in 2015, businesses spent more than twice as much as the federal government on R&D. Table 1 provides some points of international comparison for these statistics, tabulating R&D expenditures and R&D as a share of GDP in the United States, the nine other largest economies (as measured by GDP in 2015), and the OECD average. The United States spends more on R&D than these other countries, but R&D as a share of GDP in the United States is smaller than in Germany and Japan.

In recent years, around 13 percent of US research and development has been performed at colleges and universities. This R&D is also relatively unique in the sense that just under half of US R&D on basic research is undertaken at colleges and universities. From the perspective of these institutions, in recent years just over half of R&D expenditures at US colleges and universities have been federally funded. The vast bulk of that funding goes to the life sciences, with smaller amounts going to engineering, the physical sciences, and other fields.

Another set of metrics of innovative activity focus on the scientific workforce. The fraction of workers who are researchers grew through 2000 in the United States but has been stable between 0.7 and 0.9 percent since. The European Union has a similar fraction, while Japan is closer to 1 percent.

[1] Unless otherwise noted, all data and facts in this section—and later in the paper—are drawn from National Science Board (2018).

Figure 1
US Research and Development as a Share of GDP, by Source of Funds: 1953–2015



Source: This figure displays data from figure 4-3 of National Science Board (2018), chap. 4. The original data are drawn from the National Science Foundation, National Center for Science and Engineering Statistics, National Patterns of R&D Resources (annual series).
Notes: The figure shows how spending on R&D performed in the United States, presented as a share of GDP, has evolved over time from 1953 to 2015, in total and broken down by source of R&D funding.

One additional metric relevant to the size of the US scientific workforce is the number of temporary work visas issued in categories that cover high-skilled workers: J-1 (exchange visitors), H-1B, and L-1 (intracompany transferee) visas. Between 1991 and 2015, the primary increase in these categories was in J-1 visas, which increased from around 150,000 to over 330,000. The number of H-1B visas increased from around 52,000 in 1991 to nearly 175,000 in 2015. A cap of 65,000 H-1B visas was in place over that entire period, implying that the growth was driven by H-1Bs issued to employees of universities, nonprofit research facilities, and government research facilities, all of which are exempt from the annual H-1B quotas.

Why Should Governments Promote Innovation?

Governments often want to increase innovation in an attempt to encourage economic growth; indeed, countries that have higher levels of research and development spending are typically richer (see, for example, Jones 2015). However, standard economic theory suggests that, in the absence of market failures, it would be better for the government to leave investment decisions in the hands of private firms. There are many oft-cited government failures, such as the Concorde Anglo-French supersonic jet (for many other examples, see Lerner 2009). On

Table 1

International Comparison of Research and Development Expenditures in 2015

Country	*R&D expenditures (billions of US$)*	*R&D/GDP (%)*
United States	496.6	2.7
China	408.8	2.1
India	50.3	0.6
Japan	170.0	3.3
Germany	114.8	2.9
Russia	38.1	1.1
Brazil	38.4	1.2
France	60.8	2.2
United Kingdom	46.3	1.7
Indonesia	2.1	0.1
OECD (average)	34.7	2.4

Source: These data are drawn from table 4-5 of National Science Board (2018), chap. 4. The original data are drawn from the OECD, Main Science and Technology Indicators (2017/1); United Nations Educational, Scientific, and Cultural Organization Institute for Statistics Data Centre (http://data.uis.unesco.org/; accessed October 13, 2017).

Notes: This table displays data on gross domestic expenditures on R&D (reported in purchasing power parity adjusted billions of US dollars) and R&D as a share of GDP for the United States, the nine other countries with the largest GDP in 2015, and the OECD average (averaged over all 36 member countries as of 2015).

the other hand, there are also many examples of impressive inventions built on government-sponsored R&D, such as jet engines, radar, nuclear power, the Global Positioning System (GPS), and the internet (Janeway 2012; Mazzucato 2013).

Knowledge spillovers are the central market failure on which economists have focused when justifying government intervention in innovation. If one firm creates something truly innovative, this knowledge may spill over to other firms that either copy or learn from the original research—without having to pay the full research and development costs. Ideas are promiscuous; even with a well-designed intellectual property system, the benefits of new ideas are difficult to monetize in full. There is a long academic literature documenting the existence of these positive spillovers from innovations.

That said, economic theory also suggests that research and development expenditures in a market economy can be either too low or too high, depending on the net size of knowledge spillovers relative to what could be termed product market spillovers. The key idea behind product market spillovers is that private incentives can lead to business-stealing overinvestment in R&D because innovator firms may steal market share from other firms without necessarily generating any social benefit. A classic example is the case of pharmaceuticals, where one firm may spend billions of dollars to develop a drug that is only incrementally better than a drug produced by a rival firm—a “me too” drug. However, the small improvement in therapeutic value may allow the second firm to capture nearly the entire market. In cases where "me too" drugs are therapeutically indistinguishable from the products that they replace (and setting aside the possibility that such drugs may generate the benefit of price-cutting competition), this dynamic potentially generates a massive private benefit for shareholders of pharmaceutical firms, with little gain for patients.

Broadly stated, three methods have been used to estimate spillovers: case studies, a production function approach, and research based on patent counts.

Perhaps the most famous example of the case study approach is Griliches (1958), which estimates the social rate of return realized by public and private investments in hybrid corn research. Griliches estimates an annual return of 700 percent, as of 1955, on the average dollar invested in hybrid corn research. Seed or corn producers appropriated almost none of these returns; they were instead passed to consumers in the form of lower prices and higher output. While this study is widely cited, Griliches himself discusses the challenges inherent in calculating the rate of return on something akin to a successful "oil well." Although we typically observe an estimate that captures the cost of drilling and developing a successful well, we would ideally prefer to generate an estimate that includes the cost of all of the "dry holes" drilled before oil was struck. For more specific examples of diffusion, see the data compiled by Comin and Hobijn (2010).

The production function approach abandons the details of specific technologies and instead relates productivity growth (or other measures of innovative output) to lagged measures of investment in research and development. The key challenge here is that R&D is determined by many factors that also independently affect productivity. Recent papers applying this approach have used policy experiments that influence R&D investments to identify the arrow of causality (for example, Bloom, Schankerman, and Van Reenen 2013).

The key idea in using patent citations to measure spillovers is that each patent cites other patents, all of which form the basis of "prior art"—existing innovations that enabled that particular patent. Trajtenberg (1990) and Jaffe, Trajtenberg, and Henderson (1993) pioneered this approach. Although there is some evidence that citations can be strategic (and that some citations are added by patent examiners during the course of the patent examination process), the existence of patent citations provides a measurable indication of knowledge spillovers (see, for example, Griffith, Lee, and Van Reenen 2011). As already noted, a challenge with the production function approach is finding ways of identifying the relevant channels of influence so that "one can detect the path of spillovers in the sands of the data" (Griliches 1992). Herein lies an advantage of using patent citations, which provide a direct way of inferring *which* firms receive spillover benefits.

More generally, the trick in the search for spillovers has been to focus on defining a dimension (or dimensions) over which spillovers are mediated. Firms less distant from each other in this dimension will be more affected by the research and development efforts of their peers. Examples include technological distance as revealed from past patenting classes (Jaffe 1986), geographical distance between corporate R&D labs, and product market distance (the industries in which firms operate). As a whole, this literature on spillovers has consistently estimated that social returns to R&D are much higher than private returns, which provides a justification for government-supported innovation policy. In the United States, for example, recent estimates in Lucking, Bloom, and Van Reenen (2018) used three decades of firm-level data and a production function–based approach to document evidence of substantial positive net knowledge spillovers. The authors estimate that social returns are about 60 percent, compared with private returns of around 15 percent, suggesting the case for a substantial increase in public research subsidies.

Given this evidence on knowledge spillovers, one obvious solution is to provide strong intellectual property rights such as patents to inventors as a means of increasing the private return to inventing. A patent is a temporary right to exclude others from selling the protected invention. Patents entail some efficiency loss because they usually enable sellers to charge a higher price markup over production costs. However, this downside could be outweighed by the gains in dynamic efficiency that arise from patents providing stronger incentives to do more research and development because potential innovators expect to be able to appropriate more of the benefits for their efforts. In practice, as we will discuss in more detail below, the patent system is highly imperfect. For one thing, other firms can frequently invent around a patent—after all, the empirical evidence on knowledge spillovers summarized above is drawn from data on the United States, which already has a strong system of intellectual property rights by international standards.

In addition to spillovers, there are other potential justifications for research and development subsidies, related to failures in other markets. For example, financial constraints may limit the amount of innovation that firms can carry out. Because innovation is intangible, it may be hard for firms to raise funding when they have no collateral to pledge to banks in return for debt funding. This insight suggests that equity might be a better source of funding for innovation, but equity faces a different challenge: an asymmetry of information. Before innovations are patented or demonstrated in the market, the requisite secrecy about technology makes fundraising difficult. A pitch of "trust me, I have a great idea, so please fund me" is rarely effective, whereas a pitch of "let me describe my not-yet-patented idea in detail" opens up the possibility of potential investors stealing an idea from the entrepreneur.

Evidence suggests that financial constraints often do hold back innovation (for a survey, see Hall and Lerner 2010). However, the presence of financial constraints around research and development funding is not necessarily a reason for government subsidies: governments often have worse information about project quality than either firms or investors, so designing appropriate policy interventions is difficult. Effective policies to address financial constraints involve not just financial support for firms but also a mechanism to identify and select higher-quality investments acurately, which is typically difficult to do.

We now turn to discussing a number of the main innovation policy levers: tax policies to favor research and development, government research grants, policies aimed at increasing the supply of human capital focused on innovation, intellectual property policies, and pro-competitive policies.

Tax Incentives for Research and Development

The tax code automatically treats research and development expenditures by firms more generously than tangible capital investment. In particular, because most R&D expenses are current costs—like scientists' wages and lab materials—they can be written off in the year in which they occur. By contrast, investments in long-lasting assets such as plant, equipment, and buildings must be written off over a multiyear period; this allows a firm to reduce its tax liabilities only at some point in the future.

But over and above this tax structure advantage, many countries provide additional fiscal incentives for research and development, such as allowing an additional deduction to be made against tax liabilities. For example, if firms treat 100 percent of their R&D as a current expense, and the corporate income tax rate is 20 percent, then every $1 of R&D expenditure reduces corporate taxes by $0.20. However, if a government allows a 150 percent rate of superdeduction, again assuming a corporate tax rate of 20 percent, then $1 of R&D spending would reduce corporate taxes by $0.30. President Reagan introduced the first Research and Experimentation Tax Credit in the United States in 1981. This policy currently costs the US federal government about $11 billion a year in foregone tax revenue (National Science Board 2018), with an additional $2 billion a year of lost tax revenue from state-level R&D tax credits (which started in Minnesota in 1982).

The OECD (2018) reports that 33 of the 42 countries it examined provide some material level of tax generosity toward research and development. The US federal R&D tax credit is in the bottom one-third of OECD nations in terms of generosity, reducing the cost of US R&D spending by about 5 percent. This is mainly because the US tax credit is based on the incremental *increase* in a firm's R&D over a historically defined base level, rather than being a subsidy based on the total amount of R&D spending. In countries with the most generous provisions, such as France, Portugal, and Chile, the corresponding tax incentives reduce the cost of R&D by more than 30 percent.

Do research and development tax credits actually work to raise R&D spending? The answer seems to be "yes." One narrow approach to the question asks whether the quantity of R&D increases when its tax price falls. This question is of interest in part because most people (and many expert surveys) suggest that R&D is driven by advances in basic science and perhaps by market demand, rather than by tax incentives. There are now a large number of studies that examine changes in the rules determining the generosity of tax incentives by using a variety of data and methodologies (for a survey, see Becker 2015). Many early studies used cross-country panel data (Bloom, Griffith, and Van Reenen 2002) or US cross-state data (Wilson 2009) and related changes in R&D to changes in tax rules. Some more recent studies have used firm-level data and exploited differential effects of tax rules across firms before a surprise policy change. For example, firms below a size threshold may receive a more generous tax treatment, so one can compare firms just below and just above the threshold after (and before) the policy change by using a regression discontinuity design (Dechezleprêtre et al. 2016). Taking the macro and micro studies together, a reasonable overall conclusion would be that a 10 percent fall in the tax price of R&D results in at least a 10 percent increase in R&D in the long run; that is, the absolute elasticity of R&D capital with respect to its tax-adjusted user cost is unity or greater.

One concern for both research and policy is that firms may relabel existing expenditures as "research and development" to take advantage of the more generous tax breaks. Chen et al. (2019), for example, found substantial relabeling following a change in Chinese corporate tax rules. A direct way to assess the success of the R&D tax credit is to look at other outcomes such as patenting, productivity, or jobs. Encouragingly, these more direct measures also seem to increase (with a lag) following tax changes (for US evidence, see Lucking 2019 and Akcigit et al. 2018; for the United Kingdom, see Dechezleprêtre et al. 2016; for China, see Chen et al. 2019; and for Norway, see Bøler, Moxnes, and Ulltveit-Moe 2015).

Another concern is that research and development tax credits may not raise aggregate R&D but rather may simply cause a relocation toward geographical areas with more generous fiscal incentives and away from geographical areas with less generous incentives. US policymakers may not care so much if tax credits shift activity from, say, Europe to the United States, but we expect them to care if state-specific credits simply shift around activity from one state to another. There are a wide variety of local policies explicitly trying to relocate innovative activity across places within the United States by offering increasingly generous subsidies. For example, Amazon's second headquarters generated fierce competition, with some cities offering subsidies up to $5 billion. This is likely to cause some distortions, as the areas that bid the most are not always the places where the research will be most socially valuable.

There is some evidence of relocation in response to tax incentives. In the context of individual inventor mobility and personal tax rates, Moretti and Wilson (2017) find cross-state relocation within the United States, and Akcigit, Baslandze, and Stantcheva (2016) document a similar relocation pattern in an international dimension. Wilson (2009) and Bloom and Griffith (2001) also document some evidence of relocation in response to research and development tax credits. However, relocation alone does not appear to account for all of the observed changes in innovation-related outcomes. Akcigit et al. (2018) test explicitly for relocation and estimate effects of tax incentive changes on nonrelocating incumbents. Overall, the conclusion from this literature is that despite some relocation across place, the aggregate effect of R&D tax credits at the national level both on the volume of R&D and on productivity is substantial.

Patent Boxes

"Patent boxes," first introduced by Ireland in the 1970s, are special tax regimes that apply a lower tax rate to revenues linked to patents relative to other commercial revenues. By the end of 2015, patent boxes (or similarly structured tax incentives related to intellectual property) were used in 16 OECD countries (Guenther 2017). Although patent box schemes purport to be a way of incentivizing research and development, in practice they induce tax competition by encouraging firms to shift their intellectual property royalties into different tax jurisdictions. Patent boxes provide a system through which firms can manipulate stated revenues from patents to minimize their global tax burden (Griffith, Miller, and O'Connell 2014) because firms—particularly multinational firms—have considerable leeway in deciding where they will book their taxable income from intellectual property. Although it may be attractive for governments to use patent box policies to collect footloose tax revenues (Choi 2019), such policies do not have much effect on the real location or the quantity of either R&D or innovation. Gaessler, Hall, and Harhoff (2018) find a small effect of the introduction of patent boxes in several EU countries on transfers of the ownership of patents, but zero effect on real invention.

Our take is that patent boxes are an example of a harmful form of tax competition that distorts the tax system under the guise of being a pro-innovation policy. In contrast to well-designed research and development tax credits—for which it is hard to manipulate the stated location of research labs—patent boxes should be discouraged.

Government Research Grants

A disadvantage of tax-based support for research and development is that tax policies are difficult to target at the R&D that creates the most knowledge spillovers and avoids business-stealing. In contrast, government-directed grants can more naturally do this type of targeting by focusing on, for example, basic R&D, such as that performed in universities, rather than more applied R&D that occurs in an industry setting. A variety of government programs seek to encourage innovation by providing grant funding, either to academic researchers—such as through the US National Institutes of Health (NIH)—or to private firms, such as through the Small Business Innovation Research (SBIR) program. How effective are these programs?

Evaluating the effectiveness of grant funding for research and development is challenging. Public research grants usually (and understandably) attempt to target the most promising researchers, the most promising projects, or the most socially important problems. As a result, it is difficult to construct a counterfactual for what would otherwise have happened to the researchers, firms, or projects that receive public R&D funds. If $1 of public R&D simply crowds out $1 of private R&D that would otherwise have been invested in the same project, then public R&D could have no real effect on overall R&D allocations (much less on productivity or growth). However, it is also possible that public R&D grants add to private R&D spending, or even that public R&D "crowds in" and attracts additional private R&D spending.

Jacob and Lefgren (2011) use administrative data on US grant applications to the National Institutes of Health and effectively compare academic applicants who just barely received and just missed receiving large NIH grants. They document that these grants produce positive but small effects on research output, leading to about one additional publication over five years (an increase of 7 percent). One explanation for this modest effect is that marginal unsuccessful NIH grant applicants often obtain other sources of funding to continue their research. Consistent with that story, productivity effects are larger among researchers who are likely to be more reliant on NIH funding (for whom alternative funding sources may be less likely to be available).

Looking beyond academic output, public research and development grants may affect private firms in several ways. First, public R&D grants to academics can generate spillovers to private firms. Azoulay, Graff Zivin, et al. (2019) exploit quasi-experimental variation in funding from the National Institutes of Health across research areas to show that a $10 million increase in NIH funding to academics leads to 2.7 additional patents filed by private firms. Second, private firms themselves sometimes conduct publicly funded R&D. Moretti et al. (2019) use changes in military R&D spending, which is frequently driven by exogenous political changes, to look at the effect of public subsidies for military R&D. They document that a 10 percent increase in publicly funded R&D to private firms results in a 3 percent increase in private R&D, suggesting that public R&D crowds in private R&D (and also, they document, raises productivity growth). Third, private firms can directly receive public subsidies. Howell (2017) examines outcomes for Small Business Innovation Research grant applicants, comparing marginal winners and losers. She estimates that early-stage SBIR grants roughly double the probability that a firm receives subsequent venture capital funding, and that receipt of an SBIR grant has positive impacts on firm revenue and patenting.

Two other important aspects of public grant support for research and development are worth mentioning. First, a substantial share of public R&D subsidies goes to universities, which makes sense from a policy perspective, as spillovers from basic academic research are likely to be much larger than those from near-market applied research. There certainly appears to be a correlation between areas with strong science-based universities and private sector innovation (for example, Silicon Valley in California, Route 128 in Massachusetts, and the Research Triangle in North Carolina). Jaffe (1989) pioneered research in this area by documenting important effects of academic R&D on corporate patenting, a finding corroborated by Belenzon and Schankerman (2013) and Hausman (2018).[2]

Governments can also fund their own research and development labs—for example, SLAC National Accelerator Laboratory at Stanford University. These labs can generate more research activity and employment in the technological and geographical area in which the lab specializes. For example, the United Kingdom's Diamond Light Source synchrotron appeared to do this (Helmers and Overman 2016), but in that case the increase seems to have occurred mainly through relocation of research activity within the United Kingdom rather than an overall increase in aggregate research.

[2] Jaffe and Lerner (2001) analyze national labs, which are often managed by universities, and also document evidence of spillovers. Valero and Van Reenen (2019) offer a generally positive survey on the impact of universities on productivity overall and on innovation specifically. Hausman (2018) and Andrews (2019) also find positive effects of universities on US innovation.

There has also been controversy over how to design complementary policies that enable the resulting discoveries—when made at universities—to be translated into technologies that benefit consumers. The 1980 Bayh–Dole Act in the United States made some key changes in the ownership of inventions developed with public research and development support. In part because of Bayh–Dole, universities have an ownership share in the intellectual property developed by those working at their institutions, and many universities set up "technology transfer offices" to provide additional support for the commercialization of research. Lach and Schankerman (2008) provide evidence consistent with greater ownership of innovations by scientists being associated with more innovation. In addition, evidence from Norway presented in Hvide and Jones (2018) suggests that when university researchers enjoy the full rights to their innovations, they are more likely to patent inventions as well as launch start-ups. That is, ideas that might have remained in the "ivory tower" appear more likely to be turned into real products because of changes in the financial returns to academic researchers.

Human Capital Supply

So far, we have focused attention on policies that increase the demand for research and development by reducing its cost via the tax system or via direct grant funding. However, consider an example in which we assume that scientists carry out all R&D and that the total number of scientists is fixed. If the government increases demand for R&D, the result will simply be higher wages for scientists, with zero effect on the quantity of R&D or innovation. Of course, this example is extreme. There is likely to be some ability to substitute away from other factors into R&D. Similarly, there is likely some elasticity of scientist supply in the long run as wages rise and, through immigration from other countries, in the short run.[3] However, the underlying message is that increasing the quantity of innovative activity requires increasing the supply of workers with the human capital needed to carry out research, as emphasized by Romer (2001). This rise in supply increases the volume of innovation directly as well as boosting R&D indirectly by reducing the equilibrium price of R&D workers. In addition, since these workers are highly paid, increasing the supply of scientific human capital will also tend to decrease wage inequality.

Many policy tools are available that can increase the supply of scientific human capital. In terms of frontier innovation, perhaps the most direct policy is to increase the quantity and quality of inventors. There have been many attempts to increase the number of individuals with training in science, technology, engineering, and mathematics (commonly known as STEM). Evaluating the success of such policies

[3] This insight also suggests that general equilibrium effects of a research and development tax credit may partially undermine its effects on innovation. These effects are hard to detect with micro data. Some macro studies do show partial crowding out (Goolsbee 1998), whereas others do not (Bloom, Griffith, and Van Reenen 2002). Atkeson and Burstein (forthcoming) put these together in a macro model that shows large long-run welfare effects of innovation policies.

is difficult given that these policies tend to be economy-wide, with effects that will play out only in the long run.

One strand of this literature has focused on the location, expansion, and regulation of universities as key suppliers of workers in science, technology, engineering, and mathematics. For example, Toivanen and Väänänen (2016) document that individuals growing up around a technical university (such institutions rapidly expanded in the 1960s and 1970s in Finland) were more likely to become engineers and inventors. Of course, such policies could increase the supply of workers with qualifications in STEM fields, but research and innovation by university faculty could also directly affect local area outcomes.

Bianchi and Giorcelli (2018) present results from a more direct test of the former explanation by exploiting a change in the enrollment requirements for Italian majors in science, technology, engineering, and mathematics, which expanded the number of graduates. They document that this exogenous increase in STEM majors led to more innovation in general, with effects concentrated in particular in chemistry, medicine, and information technology. They also document a general "leakage" problem that may accompany efforts to simply increase the STEM pipeline: many STEM-trained graduates may choose to work in sectors that are not especially focused on research and development or innovation, such as finance.

Migration offers an alternative lens into the effects of human capital on innovation. Historically, the United States has had a relatively open immigration policy that helped to make the nation a magnet for talent. Immigrants make up 18 percent of the US labor force aged 25 and over but constitute 26 percent of the science, technology, engineering, and mathematics workforce. Immigrants also own 28 percent of higher-quality patents (as measured by those filed in patent offices of at least two countries) and hold 31 percent of all PhDs (Shambaugh, Nunn, and Portman 2017). A considerable body of research supports the idea that US immigrants, especially high-skilled immigrants, have boosted innovation. For example, Kerr and Lincoln (2010) exploit policy changes affecting the number of H1-B visas and argue that the positive effects come solely through the new migrants' own innovation.[4] Using state panel data from 1940 to 2000, Hunt and Gauthier-Loiselle (2010) document that a 1 percentage point increase in immigrant college graduates' population share increases patents per capita by 9 to 18 percent, and they argue for a spillover effect to the rest of the population. Bernstein et al. (2018) use the death of an inventor as an exogenous shock to team productivity and argue for large spillover effects of immigrants on native innovation.

The US federal government's introduction of immigration quotas with varying degrees of strictness in the early 1920s—for example, Southern Europeans, such as Italians, were more strongly affected than Northern Europeans, such as Swedes—has

[4] Using H1-B visa lotteries, Doran, Gelber, and Isen (2014) estimate smaller effects than Kerr and Lincoln (2010). By contrast, Borjas and Doran (2012) document negative effects on publications by Americans in mathematics journals following the fall of the Soviet Union, although they do not attempt to estimate aggregate effects; their findings may reflect a feature specific to academic publishing, where there are (short-run) constraints on the sizes of academic journals and departments. Moser, Voena, and Waldinger (2014) estimate that most of the effect of immigration on innovation came from new entry.

been used to document how exogenous reductions in immigration damaged innovation. Moser and San (2019) use rich biographical data to show that these quotas discouraged Eastern and Southern European scientists from coming to the United States and that this reduced aggregate invention. Doran and Yoon (2018) also find negative effects of these quotas. Moser, Voena, and Waldinger (2014) show that American innovation in chemistry was boosted by the arrival of Jewish scientists who were expelled by the German Nazi regime in the 1930s.

Overall, most of the available evidence suggests that increasing the supply of human capital through expanded university programs and/or relaxed immigration rules is likely to be an effective innovation policy.

A final way to increase the quantity supplied of research and development is to reduce the barriers to talented people becoming inventors in the first place. Children born in low-income families, women, and minorities are much less likely to become successful inventors. Bell et al. (2019), for example, document that US children born into the top 1 percent of the parental income distribution are ten times more likely to grow up to be inventors than are those born in the bottom half of the distribution. The authors show that relatively little of this difference is related to innate ability. A more important cause of the lower invention rate for disadvantaged groups appears to be differential exposure rates to inventors in childhood. This implies that improved neighborhoods, better school quality, and greater exposure to inventor role models and mentoring could arguably raise long-run innovation.

Intellectual Property

The phrase "intellectual property" is often used to refer to a suite of policies including patents, copyrights, and other instruments such as trademarks. Although these policies have some broad similarities, they differ in meaningful ways. For example, a patent grants—in exchange for disclosure of an invention—a limited-term property right to an inventor, during which time the inventor has the right to exclude others from making, using, or selling their invention. A copyright, in contrast, provides a limited term of protection to original literary, dramatic, musical, and artistic works, during which time the author has the right to determine whether, and under what conditions, others can use their work. The legal rules governing patents and copyrights are distinct, and the practical details of their implementation are quite different; for example, copyright exists from the moment a work is created (although as a practical matter it can be difficult to bring a lawsuit for infringement if you do not register the copyright), whereas an inventor must actively choose to file a patent application, and patent applications are reviewed by patent examiners. Nonetheless, patents and copyrights have many similarities from an economic perspective, and economists—to the chagrin of some lawyers—often lump the two types of policies together.

Boldrin and Levine (2013, in this journal) have argued that the patent system should be completely abolished, based on the view that there is no evidence that patents serve to increase innovation and productivity. Although the patent system has many problems, outright abolition is—in our view—an excessive response. However, many different elements of patents could be strengthened or loosened. We focus here on two specific areas currently under active policy debate.

First, what types of technologies should be patent eligible? The US Patent and Trademark Office is tasked with awarding patent rights to inventions that are novel, nonobvious, and useful and whose application satisfies the public disclosure requirement. The US Supreme Court has long interpreted Section 101 of Title 35 of the US Code as implying that abstract ideas, natural phenomena, and laws of nature are patent-ineligible. Several recent Court rulings have relied on Section 101 to argue that various types of inventions should no longer be patent eligible: business methods (*Bilski v. Kappos*, 561 US 593 [2010]), medical diagnostic tests (*Mayo Collaborative Services v. Prometheus Laboratories, Inc.*, 566 US 66 [2012]), human genes (*Association for Molecular Pathology v. Myriad Genetics, Inc.*, 569 US 576 [2013]), and software (*Alice Corp. v. CLS Bank International*, 573 US 208 [2014]). A reasonable interpretation of these legal rulings is that the Court is "carving out" certain areas where the perceived social costs of patents outweigh the perceived social benefits. For example, in the 2012 *Mayo v. Prometheus* case, the Court argued that the patenting of abstract ideas such as medical diagnostic tests might impede, more than encourage, innovation. This question is fundamentally empirical, but the available empirical evidence provides only rather inconclusive hints at the answer to that question, rather than a systematic basis for policy guidance (Williams 2013, 2017; Sampat and Williams 2019).

Second, many current debates about patent reform center on "patent trolls," a pejorative term that refers to certain "nonpracticing entities," or patent owners who do not manufacture or use a patented invention but instead buy patents and then seek to enforce patent rights against accused infringers. The key question here is whether litigation by so-called patent trolls is frivolous. On one hand, Haber and Levine (2014) argue that the recent uptick in patent litigation generally associated with the rise of patent trolls may in fact not be evidence of a problem. They argue that—historically—spikes in litigation have coincided with the introduction of disruptive technologies (such as the telegraph and the automobile) and that there is no evidence that the current patent system either harms product quality or increases prices. On the other hand, Cohen, Gurun, and Kominers (2016) find that nonpracticing entities (unlike practicing entities) sue firms that experience increases in their cash holdings. They interpret this interesting connection as evidence that—on average—nonpracticing entities act as patent trolls, but this evidence provides little information about the importance of these types of incentives in explaining the broader observed trends in patenting or innovation. While several other author teams have investigated various aspects of patent trolling (Abrams, Akcigit, and Grennan 2018; Lemley and Simcoe 2018; Feng and Jaravel forthcoming), the past literature has struggled to establish clear evidence that many or most nonpracticing entities are associated with welfare-reducing behavior.

Product Market Competition and International Trade

The impact of competition on innovation is theoretically ambiguous. On the negative side, Schumpeter (1942) argued that the desired reward for innovation is monopoly profits, and increasing competition tends to reduce those incentives. More broadly, settings with high competition may tend to imply lower future profits, which in turn will limit the internal funds available to finance research and development, which may be important given the financial frictions discussed above.

But there are also ways in which competition may encourage innovation. First, monopolists who benefit from high barriers to entry have little incentive to innovate and replace the stream of supernormal profits they already enjoy, in contrast to a new entrant who has no rents to lose (this is the "replacement effect," described in Arrow 1962). Second, tougher competition can induce managers to work harder and innovate more. Finally, capital and labor are often "trapped" within firms (for example, restricted by the costs of hiring employees or moving capital). If competition removes the market for a firm's product, it will be forced to innovate to redeploy these factors (Bloom et al. 2019). In some models, the impact of competition on innovation is plotted as an inverted U: when competition is low, the impact of greater competition on innovation first is positive, then becomes negative at higher levels of competition (see, for example, Aghion et al. 2005).

The bottom line is that the net impact of competition on innovation remains an open empirical question. However, existing empirical evidence suggests that competition typically increases innovation, especially in markets that initially have low levels of competition. Much of this literature focuses on import shocks that increase competition, such as China's integration in the global market following accession to the World Trade Organization in 2001. Shu and Steinwender (2019) summarize over 40 papers on trade and competition, arguing that in South America, Asia, and Europe, competition mostly drives increases in innovation (also see Blundell, Griffith, and Van Reenen 1999; Bloom, Draca, and Van Reenen 2016). In North America, the impact of import competition is more mixed; for example, Autor et al. (2016) argue that Chinese import competition reduced innovation in US manufacturing, although Xu and Gong (2017) argue these research and development employees displaced from manufacturing were re-employed in services, generating an ambiguous overall impact.

In addition to its effect on competition, trade openness can increase innovation by increasing market size, thus spreading the cost of innovation over a larger market (for example, Grossman and Helpman 1991). Moreover, trade leads to improved inputs and a faster diffusion of knowledge (for example, Diamond 1997; Keller 2004). Aghion et al. (2018) use shocks to a firm's export markets to demonstrate large positive effects on innovation in French firms. Atkin et al. (2017) implemented a randomized controlled trial to stimulate exports in small apparel firms in Egypt and found that exporting increases firms' productivity and quality. The benefits of superior imported inputs have been shown in a number of papers (including Goldberg et al. 2010; Fieler and Harrison 2018).

In our view, the policy prescription from this literature seems reasonably clear: greater competition and trade openness typically increase innovation. The financial costs of these policies are relatively low, given that there are additional positive impacts associated with policies that lower prices and increase choice. The downside is that such globalization shocks may increase inequality among people and places.

Targeting Small Firms

Financial constraints are often the rationale for focusing innovation policies on small firms. For example, in many countries the research and development tax credit is more generous for smaller firms (OECD 2018). Moreover, small firms appear to respond more positively to innovation and other business support policies than larger firms (Criscuolo et al. 2019). However, small-is-beautiful innovation policies have some problems as well. First, they can discourage firms from growing, as expanding beyond a certain point would disqualify them from their subsidies. Second, it is young firms, rather than small firms per se, that are most subject to these financial constraints.

One popular policy seeks to co-locate many smaller high-tech firms together. This may be in a high-density accelerator (intensive mentoring; highly selected) or incubator (less support; less selected) or in a larger science park. The idea is to generate agglomeration effects. There are several case studies and one metareview of this approach that suggest the overall impact of these policies is positive (Madaleno et al. 2018). Our sense, however, is that the evidence remains ambiguous here, despite the great popularity of these initiatives with local governments.

To the extent that financial frictions are impactful, removing constraints on the development of an active early-stage finance market (like angel finance or venture capital) might be a reasonable policy focus. In addition, focusing on subsidized loans for young firms, rather than general tax breaks or grants, may be more desirable.

More Moonshots? A Mission-Oriented Approach

Throughout this article, we have taken a pragmatic and marginal approach: given a policymaker's constraints, what is the best use of resources to stimulate growth through innovation? However, this approach may be too conservative given the scale of the current productivity problems.

Instead, some recent proposals have aimed at spurring a step change in productivity growth. Taking inspiration from the research and development efforts during World War II and Kennedy's Apollo "moonshot," "mission-oriented" R&D policies focus support on particular technologies or sectors. Many such mission-oriented policies in defense (such as DARPA, the Defense Advanced Research Projects Agency) and space (such as NASA, the National Aeronautics and Space Administration) have led to important innovations. Azoulay, Fuchs, et al. (2019) offer a detailed discussion of the "ARPA model"—an approach that has expanded beyond DARPA to HSARPA in the Department of Homeland Security, IARPA for US intelligence agencies, and ARPA-E in the Department of Energy. They argue that successful examples typically involve decentralization, active project selection (and a tolerance for inevitable failures), and organizational flexibility.

Economists are often skeptical of such sector-focused policies, because political decision-making may be more likely to favor sectors or firms that engage in lobbying and regulatory capture, rather than the most socially beneficial. Moreover, in many cases it may be hard to articulate an economic rationale behind these moonshots. Surely, the resources used in putting a man on the moon could have been directed more efficiently if the aim was solely to generate more innovation.

We see two main arguments for mission-based moonshots. First, moonshots may be justified in and of themselves. Technology to address climate change falls into this category: there is a pressing need to avoid environmental catastrophe, and obvious market failures exist around carbon emissions. The solution requires new technologies to help deliver decarbonization of the economy; moonshot strategies may result in the most valuable innovation in this case. Similar comments could be made of other social goals, such as disease reduction. It is important to remember that when the rate and direction of technological change are endogenous, conventional policies such as a carbon tax can be doubly effective (both by reducing carbon emissions and by generating incentives to direct research and development toward green technologies; see Acemoglu et al. 2012; Aghion et al. 2016).

Second, moonshots may be justified on the basis of political economy considerations. To generate significant extra resources for research, a politically sustainable vision needs to be created. For example, Gruber and Johnson (2019) argue that increasing federal funding of research as a share of GDP by half a percent—from 0.7 percent today to 1.2 percent, still lower than the almost 2 percent share observed in 1964 in Figure 1—would create a $100 billion fund that could jump-start new technology hubs in some of the more educated but less prosperous American cities (such as Rochester, New York, and Pittsburgh, Pennsylvania). They argue that such a fund could generate local spillovers and, by alleviating spatial inequality, be more politically sustainable than having research funds primarily flow to areas with highly concentrated research, such as Palo Alto, California, and Cambridge, Massachusetts.

Of course, it is difficult to bring credible econometric evidence to bear on the efficacy and efficiency of moonshots. We can discuss historical episodes and use theory to guide our thinking, but moonshots are, by nature, highly selected episodes with no obvious counterfactuals.

Conclusions

Market economies are likely to underprovide innovation, primarily due to knowledge spillovers between firms. This article has discussed the evidence on policy tools that aim to increase innovation.

We condense our (admittedly subjective) judgements into Table 2, which could be used as a toolkit for innovation policymakers. Column 1 summarizes our read of the quality of the currently available empirical evidence in terms of both the quantity

Table 2
Innovation Policy Toolkit

Policy	*Quality of evidence (1)*	*Conclusiveness of evidence (2)*	*Net benefit (3)*	*Time frame (4)*	*Effect on inequality (5)*
Direct R&D grants	Medium	Medium	💡💡	Medium run	↑
R&D tax credits	High	High	💡💡💡	Short run	↑
Patent box	Medium	Medium	Negative	NA	↑
Skilled immigration	High	High	💡💡💡	Short to medium run	↓
Universities: incentives	Medium	Low	💡	Medium run	↑
Universities: STEM supply	Medium	Medium	💡💡	Long run	↓
Trade and competition	High	Medium	💡💡💡	Medium run	↑
Intellectual property reform	Medium	Low	Unknown	Medium run	Unknown
Mission-oriented policies	Low	Low	💡	Medium run	Unknown

Source: The authors.
Notes: This is our highly subjective reading of the evidence. Column 1 reflects a mixture of the number of studies and the quality of the research design. Column 2 indicates whether the existing evidence delivers any firm policy conclusions. Column 3 is our assessment of the magnitude of the benefits minus the costs (assuming these are positive). Column 4 delineates whether the main benefits (if there are any) are likely to be seen in the short run (roughly, the next three to four years) or in the longer run (roughly ten years or more); NA means not applicable. Column 5 lists the likely effect on inequality.

of papers and the credibility of the evidence provided by those studies. Column 2 summarizes the conclusiveness of the evidence for policy. Column 3 scores the overall benefits minus costs (that is, the net benefit), in terms of a light bulb ranking where three is the highest. This ranking is meant to represent a composite of the strength of the evidence and the magnitude of average effects. Columns 4 and 5 are two other criteria: first, whether the main effects would be short term (say, within the next three to four years), medium term, or long term (approximately ten years or more), and second, the likely effects on inequality. Different policymakers (and citizens) will assign different weights to these criteria.

In the short run, research and development tax credits and direct public funding seem the most effective, whereas increasing the supply of human capital (for example, through expanding university admissions in the areas of science, technology, engineering, and mathematics) is more effective in the long run. Encouraging skilled immigration has big effects even in the short run. Competition and open trade policies probably have benefits that are more modest for innovation, but they are cheap in financial terms and so also score highly. One difference is that R&D subsidies and open trade policies are likely to increase inequality, partly by increasing the demand for highly skilled labor and partly, in the case of trade, because some communities will endure the pain of trade adjustment and job loss. In contrast, increasing the supply of highly skilled labor is likely to reduce inequality by easing competition for scarce human capital.

Of course, others will undoubtedly take different views on the policies listed in Table 2. Nevertheless, we hope that this framework at least prompts additional debate over what needs to be done to restore equitable growth in the modern economy.

■ *For helpful comments on earlier drafts, we thank the editors (Timothy Taylor, Enrico Moretti, and Gordon Hanson) as well as Bronwyn Hall, Ben Jones, Josh Lerner, Petra Moser, and Jim Poterba. Research reported in this publication was supported in part by the Sloan Foundation, Smith Richardson Foundation, Schmitt Sciences, the Economic and Social Research Council, and the National Institute on Aging and the National Institutes of Health Common Fund, Office of the NIH Director, through grant U01-AG046708 to the National Bureau of Economic Research; the content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the National Bureau of Economic Research.*

References

Abrams, David, Ufuk Akcigit, and Jillian Grennan. 2018. "Patent Value and Citations: Creative Destruction or Strategic Disruption?" University of Pennsylvania, Institute for Law and Economics Research Paper 13-23.

Acemoglu, Daron, Philippe Aghion, Leonardo Bursztyn, and David Hemous. 2012. "The Environment and Directed Technical Change." *American Economic Review* 102(1): 131–66.

Aghion, Philippe, Antonin Bergeaud, Matthieu Lequien, and Marc J. Melitz. 2018. "The Impact of Exports on Innovation: Theory and Evidence." NBER Working Paper 24600.

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. 2005. "Competition and Innovation: An Inverted-U Relationship?" *Quarterly Journal of Economics* 120(2): 701–28.

Aghion, Philippe, Antoine Dechezleprêtre, David Hémous, Ralf Martin, and John Van Reenen. 2016. "Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry." *Journal of Political Economy* 124(1): 1–51.

Akcigit, Ufuk, Salomé Baslandze, and Stefanie Stantcheva. 2016. "Taxation and the International Mobility of Inventors." *American Economic Review* 106(10): 2930–81.

Akcigit, Ufuk, John Grigsby, Tom Nicholas, and Stefanie Stantcheva. 2018. "Taxation and Innovation in the 20th Century." NBER Working Paper 24982.

Andrews, Michael. 2019. "How Do Institutions of Higher Education Affect Local Invention? Evidence from the Establishment of U.S. Colleges." https://ssrn.com/abstract=3072565.

Arrow, Kenneth. 1962. "Economic Welfare and Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity: Economic and Social Factors*, 609–26. Princeton, NJ: Princeton University Press.

Atkeson, Andrew, and Ariel Burstein. Forthcoming. "Aggregate Implications of Innovation Policy." *Journal of Political Economy*.

Atkin, David, Amit K. Khandelwal, and Adam Osman. 2017. "Exporting and Firm Performance: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 132(2): 551–615.

Autor, David, David Dorn, Gordon H. Hanson, Gary Pisano, and Pian Shu. 2016. "Foreign Competition and Domestic Innovation: Evidence from U.S. Patents." NBER Working Paper 22879.

Azoulay, Pierre, Erica Fuchs, Anna P. Goldstein, and Michael Kearney. 2019. "Funding Breakthrough Research: Promises and Challenges of the 'ARPA Model,'" in *Innovation Policy and the Economy*, vol. 19, edited by Joshua Lerner and Scott Stern, 69–96. Chicago: University of Chicago Press.

Azoulay, Pierre, Joshua S. Graff Zivin, Danielle Li, and Bhaven N. Sampat. 2019. "Public R&D Investments and Private-Sector Patenting: Evidence from NIH Funding Rules." *Review of Economic Studies* 86(1): 117–52.

Becker, Bettina. 2015. "Public R&D Policies and Private R&D Investment: A Survey of the Empirical Evidence." *Journal of Economic Surveys* 29(5): 917–42.

Belenzon, Sharon, and Mark Schankerman. 2013. "Spreading the Word: Geography, Policy, and Knowledge Spillovers." *Review of Economics and Statistics* 95(3): 884–903.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. 2019. "Who Becomes an Inventor in America? The Importance of Exposure to Innovation." *Quarterly Journal of Economics* 134(2): 647–713.

Bernstein, Shai, Rebecca Diamond, Timothy James McQuade, and Beatriz Pousada. 2018. "The Contribution of High-Skilled Immigrants to Innovation in the United States." Stanford Graduate School of Business Working Paper 3748.

Bianchi, Nicola, and Michela Giorcelli. 2018. "Reconstruction Aid, Public Infrastructure, and Economic Development." https://ssrn.com/abstract=3153139.

Bloom, Nicholas, Mirko Draca, and John Van Reenen. 2016. "Trade Induced Technical Change? The Impact of Chinese Imports on Innovation, IT and Productivity." *Review of Economic Studies* 83(1) 87–117.

Bloom, Nicholas, and Rachel Griffith. 2001. "The Internationalisation of UK R&D." *Fiscal Studies* 22(3): 337–55.

Bloom, Nicholas, Paul Romer, Stephen J. Terry, and John Van Reenen. 2019. "Trapped Factors and China's Impact on Global Growth." http://people.bu.edu/stephent/files/TF_MAIN_DOC.pdf.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen. 2013. "Identifying Technology Spillovers and Product Market Rivalry." *Econometrica* 81(4): 1347–93.

Bloom, Nick, Rachel Griffith, and John Van Reenen. 2002. "Do R&D Tax Credits Work? Evidence from a Panel of Countries 1979–1997." *Journal of Public Economics* 85(1): 1–31.

Blundell, Richard, Rachel Griffith, and John Van Reenen. 1999. "Market Share, Market Value and Innovation: Evidence from British Manufacturing Firms." *Review of Economic Studies* 66(3): 529–54.

Boldrin, Michele, and David K. Levine. 2013. "The Case against Patents." *Journal of Economic Perspectives* 27(1): 3–22.

Bøler, Esther Ann, Andreas Moxnes, and Karen Helene Ulltveit-Moe. 2015. "R&D, International Sourcing, and the Joint Impact on Firm Performance." *American Economic Review* 105(12):3704–39.

Borjas, George J., and Kirk B. Doran. 2012. "The Collapse of the Soviet Union and the Productivity of American Mathematicians." *Quarterly Journal of Economics* 127(3): 1143–203.

Chen, Zhao, Zhikuo Liu, Juan Carlos Suárez-Serrato, and Daniel Yi Xu. 2019. "Notching R&D Investment with Corporate Income Tax Cuts in China." NBER Working Paper 24749.

Choi, Jane. 2019. "How Much Do Tax Rates Matter for the Location of Intellectual Property." Unpublished.

Cohen, Lauren, Umit Gurun, and Scott Duke Kominers. 2016. "Shielded Innovation." https://ssrn.com/abstract=2758841.

Comin, Diego, and Bart Hobijn. 2010. "An Exploration of Technology Diffusion." *American Economic Review* 100(5): 2031–59.

Criscuolo, Chiara, Ralf Martin, Henry G. Overman, and John Van Reenen. 2019. "Some Causal Effects of an Industrial Policy." *American Economic Review* 109(1): 48–85.

Dechezleprêtre, Antoine, Elias Einiö, Ralf Martin, Kieu-Trang Nguyen, and John Van Reenen. 2016. "Do Fiscal Incentives Increase Innovation? An RD Design for R&D." Centre for Economic Performance Discussion Paper 1413.

Diamond, Jared. 1997. *Guns, Germs, and Steel.* New York: W. W. Norton.

Doran, Kirk, Alexander Gelber, and Adam Isen. 2014. "The Effects of High-Skilled Immigration Policy on Firms: Evidence from H-1B Visa Lotteries." NBER Working Paper 20668.

Doran, Kirk, and Chungeun Yoon. 2018. "Immigration and Invention: Evidence from the Quota Acts." https://www3.nd.edu/~kdoran/Doran_Quotas.pdf.

Feng, Josh, and Xavier Jaravel. Forthcoming. "Crafting Intellectual Property Rights: Implications for Patent Assertion Entities, Litigation, and Innovation." *American Economic Journal: Applied Economics.*

Fieler, Ana Cecília, and Ann Harrison. 2018. "Escaping Import Competition and Downstream Tariffs." NBER Working Paper 24527.

Gaessler, Fabian, Bronwyn H. Hall, and Dietmar Harhoff. 2018. "Should There Be Lower Taxes on Patent Income?" Institute for Fiscal Studies Working Paper W18/19.

Goldberg, Pinelopi Koujianou, Amit Kumar Khandelwal, Nina Pavcnik, and Petia Topalova. 2010. "Imported Intermediate Inputs and Domestic Product Growth: Evidence from India." *Quarterly Journal of Economics* 125(4): 1727–67.

Goolsbee, Austan. 1998. "Does Government R&D Policy Mainly Benefit Scientists and Engineers?" *American Economic Review* 88(2): 298–302.

Griffith, Rachel, Sokbae Lee, and John Van Reenen. 2011. "Is Distance Dying at Last? Falling Home Bias in Fixed-Effects Models of Patent Citations." *Quantitative Economics* 2(2): 211–49.

Griffith, Rachel, Helen Miller, and Martin O'Connell. 2014. "Ownership of Intellectual Property and Corporate Taxation." *Journal of Public Economics* 112(1): 12–23.

Griliches, Zvi. 1958. "Research Costs and Social Returns: Hybrid Corn and Related Innovations." *Journal of Political Economy* 66(5): 419–31.

Griliches, Zvi. 1992. "The Search for R&D Spillovers." *Scandinavian Journal of Economics*

94(Supplement): S29–47.

Grossman, Gene M., and Elhanan Helpman. 1991. *Innovation and Growth in the Global Economy.* Cambridge, MA: MIT Press.

Gruber, Jonathan, and Simon Johnson. 2019. *Jump-Starting America: How Breakthrough Science Can Revive Economic Growth and the American Dream.* New York: Public Affairs Books.

Guenther, Gary. 2017. "Patent Boxes: A Primer." Congressional Research Service Report.

Haber, Stephen, and Ross Levine. 2014. "The Myth of the Wicked Patent Troll." *Wall Street Journal*, June 29, 2014. https://www.wsj.com/articles/stephen-haber-and-ross-levine-the-myth-of-the-wicked-patent-troll-1404085391.

Hall, Bronwyn H., and Josh Lerner. 2010. "The Financing of R&D and Innovation," in *Handbook of the Economics of Innovation*, vol. 1, edited by Bronwyn H. Hall and Nathan Rosenberg, 609–39. Amsterdam: Elsevier.

Hausman, Naomi. 2018. "University Innovation and Local Economic Growth." https://drive.google.com/file/d/1dfYTyG2zzVvfbRBJQ7arOLt7d_7M29Fv/view.

Helmers, Christian, and Henry G. Overman. 2016. "My Precious! The Location and Diffusion of Scientific Research: Evidence from the Synchrotron Diamond Light Source." *Economic Journal* 127(604): 2006–40.

Howell, Sabrina T. 2017. "Financing Innovation: Evidence from R&D Grants." *American Economic Review* 107(4): 1136–64.

Hunt, Jennifer, and Marjolaine Gauthier-Loiselle. 2010. "How Much Does Immigration Boost Innovation?" *American Economic Journal: Macroeconomics* 2(2): 31–56.

Hvide, Hans K., and Benjamin F. Jones. 2018. "University Innovation and the Professor's Privilege." *American Economic Review* 108(7): 1860–98.

Jacob, Brian A., and Lars Lefgren. 2011. "The Impact of Research Grant Funding on Scientific Productivity." *Journal of Public Economics* 95(9–10): 1168–77.

Jaffe, Adam B. 1986. "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value." *American Economic Review* 76(5): 984–1001.

Jaffe, Adam B. 1989. "Real Effects of Academic Research." *American Economic Review* 79(5): 957–70.

Jaffe, Adam B., and Josh Lerner. 2001. "Reinventing Public R&D: Patent Policy and the Commercialization of National Laboratory Technologies." *RAND Journal of Economics* 32(1): 167–98.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Quarterly Journal of Economics* 108(3): 577–98.

Janeway, William H. 2012. *Doing Capitalism in the Innovation Economy: Markets, Speculation and the State.* Cambridge: Cambridge University Press.

Jones, Charles I. 2015. "The Facts of Economic Growth," in *Handbook of Macroeconomics*, vol. 2A, edited by John B. Taylor and Harald Uhlig, 3–69. Amsterdam: Elsevier.

Keller, Wolfgang. 2004. "International Technology Diffusion." *Journal of Economic Literature* 42(3): 752–82.

Kerr, William R., and William F. Lincoln. 2010. "The Supply Side of Innovation: H-1B Visa Reforms and U.S. Ethnic Invention." *Journal of Labor Economics* 28(3): 473–508.

Lach, Saul, and Mark Schankerman. 2008. "Incentives and Invention in Universities." *RAND Journal of Economics* 39(2): 403–33.

Lemley, Mark A., and Timothy Simcoe. 2018. "How Essential Are Standard-Essential Patents?" Stanford Public Law Working Paper.

Lerner, Josh. 2009. *The Boulevard of Broken Dreams: Why Public Efforts to Boost Entrepreneurship and Venture Capital Have Failed—and What to Do about It.* Princeton, NJ: Princeton University Press.

Li, Xing, Megan MacGarvie, and Petra Moser. 2018. "Dead Poets' Property—How Does Copyright Influence Price?" *RAND Journal of Economics* 49(1): 181–205.

Lucking, Brian. 2019. "Do R&D Tax Credits Create Jobs?" http://stanford.edu/~blucking/jmp.pdf.

Lucking, Brian, Nicholas Bloom, and John Van Reenen. 2018. "Have R&D Spillovers Changed?" NBER Working Paper 24622.

Madaleno, Margarida, Max Nathan, Henry Overman, and Sevrin Waights. 2018. "Incubators, Accelerators and Regional Economic Development." Centre for Economic Performance Discussion Paper 1575.

Mazzucato, Mariana. 2013. *The Entrepreneurial State: Debunking Public vs. Private Sector Myths.* London: Anthem Press.

Moretti, Enrico, Claudia Steinwender, John Van Reenen, and Patrick Warren. 2019. "The Intellectual Spoils of War? Defense R&D, Productivity and International Technology Spillovers." Unpublished.

Moretti, Enrico, and Daniel J. Wilson. 2017. "The Effect of State Taxes on the Geographical Location of Top Earners: Evidence from Star Scientists." *American Economic Review* 107(7): 1858–903.

Moser, Petra, and Shmuel San. 2019. "Immigration, Science, and Invention: Evidence from the

1920s Quota Acts." Unpublished.

Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2014. "German Jewish Émigrés and US Invention." *American Economic Review* 104(10): 3222–55.

National Science Board. 2018. *Science and Engineering Indicators 2018*. Alexandria, VA: National Science Board.

OECD. 2018. *OECD Review of National R&D Tax Incentives and Estimates of R&D Tax Subsidy Rates, 2017*. Paris: OECD.

Romer, Paul M. 2001. "Should the Government Subsidize Supply or Demand in the Market for Scientists and Engineers?" Chap. 7 in *Innovation Policy and the Economy*, vol. 1, edited by Adam B. Jaffe, Josh Lerner, and Scott Stern, 221–52. Cambridge, MA: MIT Press.

Sampat, Bhaven, and Heidi L. Williams. 2019. "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome." *American Economic Review* 109(1): 203–36.

Schumpeter, Joseph A. 1942. *Capitalism, Socialism and Democracy*. New York: Harper and Brothers.

Shambaugh, Jay, Ryan Nunn, and Becca Portman. 2017. "Eleven Facts about Innovation and Patents." Hamilton Project Report, Brookings Institution.

Shu, Pian, and Claudia Steinwender. 2019. "The Impact of Trade Liberalization on Firm Productivity and Innovation," in *Innovation Policy and the Economy*, vol. 19, edited by Josh Lerner and Scott Stern, 39–68. Chicago: University of Chicago Press.

Toivanen, Otto, and Lotta Väänänen. 2016. "Education and Invention." *Review of Economics and Statistics* 98(2): 382–96.

Trajtenberg, Manuel. 1990. "A Penny for Your Quotes: Patent Citations and the Value of Innovations." *RAND Journal of Economics* 21(1): 172–87.

Valero, Anna, and John Van Reenen. 2019. "The Economic Impact of Universities: Evidence from Across the Globe." *Economics of Education Review* 68(1): 53–67.

Williams, Heidi L. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *Journal of Political Economy* 121(1): 1–27.

Williams, Heidi L. 2017. "How Do Patents Affect Research Investments?" *Annual Review of Economics* 9(1): 441–69.

Wilson, Daniel J. 2009. "Beggar Thy Neighbor? The In-State, Out-of-State, and Aggregate Effects of R&D Tax Credits." *Review of Economics and Statistics* 91(2): 431–36.

Xu, Rui, and Kaiji Gong. 2017. "Does Import Competition Induce R&D Reallocation? Evidence from the U.S." International Monetary Fund Working Paper 17/253.

How Prevalent Is Downward Rigidity in Nominal Wages? International Evidence from Payroll Records and Pay Slips

Michael W. L. Elsby and Gary Solon

In Chapter 2 of *The General Theory of Employment, Interest and Money* (1936), John Maynard Keynes put forward an assumption of downward rigidity in nominal wages as the cornerstone of his analysis of what happens in the labor market during the business cycle. According to this analysis, if the real value of the existing nominal wage exceeds the market-clearing level, downward nominal rigidity prevents arbitrage toward that level. Instead, employment is determined by the demand side of the labor market, and the excess supply of labor at that wage manifests as high unemployment. Keynes's brief theoretical account of why workers refuse to accept a nominal wage reduction, even when unemployment is the consequence, involved workers' concern about their wages relative to their reference group. Keynes did not directly address why workers would be *so* preoccupied with their relative wage that they would prefer losing their job, even during a recession, to accepting a wage cut. Keynes's empirical basis for his assumption was that, "whether logical or illogical, experience shows that this is how labour in fact behaves." He did not provide any quantitative evidence to support this observation.

In the 80-plus years since publication of *The General Theory*, Keynes's premise of downward nominal wage rigidity has continued to be highly influential. This has much to do with its potential to address some enduring macroeconomic questions: to the extent that downward rigidity prevents the real value of nominal wages from adjusting downward sufficiently in times of recession, it offers a potential account

■ *Michael W. L. Elsby is Professor of Economics, University of Edinburgh, Edinburgh, United Kingdom. Gary Solon is Professor Emeritus of Economics, University of Michigan, Ann Arbor, Michigan. Their email addresses are mike.elsby@ed.ac.uk and gary.r.solon@gmail.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.185 doi=10.1257/jep.33.3.185

for cyclical unemployment fluctuations. In addition, by implying that higher inflation might enable real wage reductions that otherwise would be impeded by downward nominal wage rigidity, it provides a potential foundation for a Phillips curve trade-off between inflation and unemployment. A quintessential implication, noted prominently in Tobin's (1972) presidential address to the American Economic Association and extended in Akerlof, Dickens, and Perry's (1996) influential paper, is that positive inflation can "grease the wheels of the labor market."

As rates of inflation have subsided in recent decades, and with the onset of the Great Recession, interest in Keynes's hypothesis of downward nominal wage rigidity has naturally revived, inspiring an array of modern applications. Formal theories of the Phillips curve in the short and long run have been developed and extended to analyze the persistent rise in US unemployment that accompanied the Great Recession (Benigno and Ricci 2011; Daly and Hobijn 2014). In international macroeconomics, the adverse interaction of downward nominal wage rigidity with currency pegs has been advanced as a key determinant of recent rises in unemployment in the eurozone and its periphery (Schmitt-Grohé and Uribe 2016). Most recently, the asymmetric nature of downward nominal wage rigidity has been invoked to provide a potential explanation for asymmetries in unemployment fluctuations over the business cycle (Dupraz, Nakamura, and Steinsson 2018).

An attractive feature of Keynes's hypothesis is that, in principle, it is amenable to empirical testing. An economy subject to a binding downward constraint on nominal wage changes should bear two hallmarks: a scarcity of nominal wage cuts and a consequent abundance of nominal wage freezes. Accordingly, a large empirical literature has sought to provide measures of the frequencies of nominal wage cuts and freezes, aided by the increasing availability of the requisite longitudinal data on individual wages.

Until recently, most such evidence had been based on reports of job stayers obtained from household surveys. That evidence, defying simple conclusions, seemed to suggest not only that nominal wage cuts are quite common (indicating a degree of downward flexibility in nominal wages) but also that nominal wage freezes are similarly common (indicating a degree of nominal rigidity). To complicate matters further, both results have been discounted on the grounds that they could be artifacts of the considerable response error in household surveys. Thus, despite the seeming testability of Keynes's hypothesis, a clear assessment of the empirical basis for downward nominal rigidity has proved elusive because of the difficulty of obtaining reliable estimates of the incidence of nominal wage changes.

The main point of the present paper is to draw attention to a more recent literature that, cumulatively, has made considerable progress on these challenges. In our view, the most compelling way to address a concern over measurement error is to seek more accurate data. The literature we survey focuses on wage data taken from employers' payroll records and pay slips. We believe this growing body of evidence has been undernoticed, perhaps because the studies have been scattered across many countries and across journals in multiple fields in economics, but also because several sources of such data have become available only recently.

Here we gather studies for Great Britain, the United States, West Germany, Austria, Italy, Spain, Mexico, Ireland, South Korea, Portugal, Sweden, and Finland. Collectively, they make an important point: except in extreme circumstances (when nominal wage cuts are either legally prohibited or rendered beside the point by very high inflation), nominal wage cuts from one year to the next appear quite common, typically affecting 15–25 percent of job stayers in periods of low inflation. Consistent with this picture of downward flexibility, nominal wage freezes are found to be much less frequent, typically affecting less than 8 percent of job stayers, and there is little evidence for large accumulations of wage freezes in times of low inflation.

None of this denies the existence of some nominal wage stickiness. Like most of our readers, we have our salaries set in nominal terms and typically see them adjusted only once a year. But does it follow from such apparent wage stickiness that nominal wages *cannot* be cut, even when inefficient layoffs or hiring decisions are the alternative? In light of the emerging evidence from more accurate wage data, we will conclude that the assumption that nominal wages cannot be cut needs to be reconsidered.

Some Modern Perspectives on the Economics of Downward Wage Rigidity

It has been more than 80 years since Keynes posited that nominal wages cannot be cut and that inefficient layoffs into unemployment are the result. As documented above, many (though far from all) modern macroeconomists still use these assumptions as key elements of their analysis. Even so, much has changed over these 80-plus years in how labor economists and macroeconomists think about the labor market, and some of the new ideas matter for the economics of downward wage rigidity and its potential effects on labor market allocations.

To begin with, the interpretation of Keynes summarized in our opening paragraph provides a simple "spot market" view of the labor market. But a distinctive characteristic of employment relationships is that they are frequently long term in nature: employees often work for the same employer for extended periods of time. This observation has important implications for the economics of wage rigidity. As noted since the seminal work of Becker (1962), the effective price of labor ceases to be simply the flow wage; rather, it is the expected present discounted value of the stream of wages anticipated over the course of the employment relationship. In addition, the seeming paradox of Keynes's theory—that workers will refuse nominal wage cuts, even when unemployment is the alternative—is thrown into sharper relief once the durability of employment relationships is acknowledged. The theory implies that an existing gainful exchange of labor is forfeit by a refusal to countenance a wage cut, even when it is mutually advantageous for both firm and worker to do so (Barro 1977). A corollary of these implications is that all that is required to obviate such inefficient layoffs is that (the present value of) wages be sufficiently flexible *at the point when separation is potentially at issue*. Subject to this requirement, flow wages can otherwise be arbitrarily rigid and indeed can accommodate many of the outward signs of downward nominal wage rigidity. Nominal wages can remain constant for periods of time if neither firm nor worker wishes to separate. And when nominal wages are adjusted, they naturally will rise more often than they fall, owing, for example, to the presence of inflation (Malcomson 1997).

This perspective cautions against leaping from the premise of apparent wage stickiness to the conclusion that inefficient layoffs, and therefore increased unemployment, must ensue. Because these arguments have informed the majority of modern macroeconomic analyses of labor markets, it is important to articulate potential channels through which rigid wages in general, and downwardly rigid nominal wages in particular, still may affect labor market allocations.

A first channel relates to another conceptual development in macroeconomic modeling that considers the implications of wage rigidity for hiring as well as layoff decisions. Becker's (1962) insight suggests that hiring incentives will be shaped by the present value of wages firms must offer to newly hired workers. Hires will fall more precipitously during recessions if firms perceive such present values to be inflexible—for example, if the wages of *both* newly hired *and* incumbent workers are sticky (Shimer 2004; Hall 2005). Importantly, there is evidence to suggest that the wages of newly hired and incumbent workers are not set in isolation. Bewley's (1999) interviews of managers highlighted the role of the internal wage structure within firms in linking the wages of new hires to those of incumbent workers. If new hires are paid according to existing wage structures, perhaps for reasons of equal treatment, any rigidity in incumbent wages is then propagated onto the wages of new hires (Gertler and Trigari 2009; Snell and Thomas 2010). An implication of this view is that any downward rigidity in nominal wages of job stayers will additionally contribute to downward nominal rigidity among new hires' wages, thereby depressing hiring incentives in times of recession.

A second channel relates to an even better-known message from Bewley's (1999) book. In a variation on Keynes's assumption that workers refuse wage cuts, what Bewley heard from the managers he interviewed was that even if they did not withdraw their labor altogether, workers disgruntled by a wage cut would be likely to exert less effort on the job. Employers therefore are reluctant to impose wage cuts for fear of adverse productivity consequences. This evidence reinforces our impression that downward wage stickiness is indeed a fact of labor market life. It is also natural to hypothesize that the prospect of such productivity losses might have allocational effects. The evidence provides a potential motive for excess layoffs that, in the words of one of Bewley's interviewees, "get the misery out the door."[1] Likewise, the anticipation of

[1] Another of Bewley's (1999) messages, which we believe the economics profession has mostly overlooked, suggests that downward stickiness may not be so extreme as to force inefficient layoff or hiring decisions. On p. 16 of his introductory chapter, Bewley says that his "mistaken" prior view had been that "an individual firm could save a significant number of jobs by reducing pay. This is seldom true, and the firms for which it is true are precisely the ones most likely to cut pay." His detailed evidence appears in his section 11.3, which begins, "I was surprised to learn that most managers did not believe that pay cuts would prevent many layoffs." This finding is altogether consistent with the Becker–Barro–Malcomson point that short-term wage stickiness need not induce inefficient allocation decisions.

such productivity losses in the future might in turn further retard firms' incentives to hire. Both of these forces might be expected to contribute to unemployment in times of recession.

Collectively, these developments in economic thinking (along with many others not discussed here) recognize that the labor market is much more complex than the bare-bones model presented in Keynes (1936). Nevertheless, there remain important potential channels through which downward wage rigidity can have unemployment consequences, on both hiring and layoff margins. We still are left with the same fundamental questions: Just how prevalent is downward rigidity in nominal wages, and what are the ramifications for the efficiency of layoff and hiring decisions? Our answers to these questions should be informed by the best available evidence, which is the subject of the remainder of this paper.

Evidence from Employer Payroll Records and Pay Slips

Most studies on nominal wage rigidity have sought to provide measures of year-to-year changes in individual workers' nominal wages from longitudinal microdata. Because much evidence shows that those changing employers typically realize wage changes, these studies have focused on the subsample of individuals who are job stayers.[2] For a long time, the majority of such measures were based on longitudinal analyses of household surveys, inspired by influential early studies of the Panel Study of Income Dynamics and the Current Population Survey in the United States (McLaughlin 1994; Card and Hyslop 1996; Kahn 1997). As we have noted, such studies typically have found not only a substantial fraction of nominal wage cuts among job stayers but also a similarly common incidence of nominal wage freezes. For example, our own 2016 *Journal of Labor Economics* paper with Donggyun Shin, which tracked job stayers from one January to the next in the Current Population Survey, found that the percentage measured as receiving a nominal wage cut was regularly between 15 and 25 percent (Elsby, Shin, and Solon 2016). In the same data, the percentage recorded with zero nominal wage change was frequently in the range of 10 to 20 percent.

However, such findings have been open to the criticism that household survey reports of wages are notoriously subject to response error. As many authors have pointed out, such errors could bias the results in either direction—that is, toward finding either more or less wage rigidity. On one hand, differences in individual response errors across survey years may exaggerate the appearance of wage flexibility: for example, someone whose nominal wage did not really decrease could still be measured as receiving a wage cut, and cases in which nominal wages truly

[2] As foreshadowed by the discussion in our previous section, an important example of what these studies have *not* attempted to measure is the rigidity of the wages of newly hired workers. Addressing this question empirically is surprisingly difficult because it calls for hiring wage data over time for the same jobs within the same firms, and such data are hard to come by. The effort by Martins, Solon, and Thomas (2012) uses the same Portuguese census of employers we cite later in this article and finds that real hiring wages in Portugal were highly procyclical over the period from 1982 to 2008.

did not change could be recorded as wage changes. Such concerns have motivated some authors, such as Akerlof, Dickens, and Perry (1996) and Altonji and Devereux (1999), to suggest that the appearance of frequent nominal wage cuts in household surveys is an artifact of measurement error. On the other hand, if wage reports are subject to rounding errors, modest wage changes will be recorded as wage freezes, exaggerating the appearance of wage rigidity. The upshot, of course, is that the nature of the bias depends on the presumed structure of response errors. Indeed, one approach taken in a portion of the literature, exemplified by some of the work discussed in this journal by Dickens et al. (2007), has attempted to correct for measurement error by imposing assumptions about the measurement error process.

The studies we review here take a more direct, and we think more persuasive, approach to addressing concerns over measurement error—namely, to seek more accurate data. In particular, we turn to administrative data from payroll records and pay slips that allow a researcher to track individual workers and the jobs they do across years and that contain accurate information on wages. Our survey identified 13 such sources of data for 12 countries. We distill relevant information from these in Table 1. For each study, the table summarizes the data source, the wage measure,[3] and the percentages of job stayers recorded as receiving either nominal wage cuts or zero change in their nominal wages. In the remainder of this section, we provide some context for the contents of Table 1. We pay particular attention to how each study addresses the measurement challenges noted above and the implications for the prevalence of downward nominal wage rigidity.

Great Britain

The first steps in the quest for more accurate wage data were taken in the British literature, so we will begin there. The first row of Table 1 summarizes the pioneering study by Smith (2000), who analyzed the 1991–1996 waves of the British Household Panel Study. In many respects, this longitudinal household survey resembles the Panel Study of Income Dynamics for the United States. Indeed, Smith's initial results based on these data resembled those based on US household surveys, measuring nontrivial minorities of respondents as receiving both wage cuts and wage freezes.

Smith also discovered, however, that the British Household Panel Study incorporated a feature that was unique at the time: respondents were allowed to check their pay slips when reporting their wages, and the survey recorded who did so. Smith's results thus provided a first glimpse of the implications of more accurate wage data for the prevalence of downward nominal wage rigidity.

The results were striking. Even among the subsample of respondents who consulted their pay slips, the incidence of nominal wage cuts remained considerable;

[3] In most instances, the measure does not include nonwage compensation. In the United States, where fringe benefits such as employer-provided health insurance loom large, this is a potentially significant omission. Lebow, Saks, and Wilson (2003) have argued that fringe benefits are an additional dimension for adjustment in compensation, so overlooking them is likely to make total compensation seem less flexible than it actually is. A similar point applies to variation in work effort.

the percentage with negative nominal wage change was 17.8 percent. By contrast, a much smaller percentage of the subsample who consulted their paychecks, just 5.6 percent, reported zero nominal wage change. Set in a context of low inflation rates—which averaged around 3 percent in Britain over Smith's sample period—the abundance of wage cuts and paucity of wage freezes are especially notable.

At the time, Smith (2000) was at pains to acknowledge surprise at her results: "Some of the results in this paper may seem difficult to believe—the quite common occurrence of nominal pay cuts, for example. It may well be that the difficulty in believing them stems not from the weight of contradictory evidence, but rather from conventional wisdom that has survived because of the previous lack of evidence either way." Since then, however, evidence amassed from a diverse range of sources has vindicated Smith's early findings.

Inspired by Smith's (2000) results, Nickell and Quintini (2003) identified another source of accurate wage data in the New Earnings Survey for Great Britain. This survey comprises a 1 percent sample of income tax–paying workers, defined by those whose National Insurance numbers (for social security) end in a given pair of digits. Because the same pair of digits has been used since the survey's inception, this survey allows one to track the same individuals over time. In the spirit of Smith's use of reports from pay slips, the New Earnings Survey data are also thought to provide unusually accurate information on individual earnings because the survey is administered to employers, who are legally required to report such information from their payroll records for a reference week each April.

The data from the New Earnings Survey also come with additional methodological advantages over the British Household Panel Study. Accompanying the data on weekly earnings are employer-reported payroll data on employee work hours for the survey reference week, permitting an analysis of hourly wages. Moreover, the New Earnings Survey records separate measures of components of earnings and hours, most notably those attributable to overtime. Because it is not obvious that, for example, reductions in hourly earnings associated with reductions in overtime should be interpreted as wage cuts, an advantage of the New Earnings Survey is that it allows one to focus on hourly wages exclusive of overtime. Finally, because it is based on a 1 percent sample of income tax–paying workers in Britain, the sample sizes it offers are large.

Nickell and Quintini's (2003) results dovetail with Smith's (2000) earlier findings. For the 1991–1996 period, over which the two studies overlap, the New Earnings Survey data produce results that mirror closely those for the respondents to the British Household Panel Study who checked their pay slips. When Nickell and Quintini widened their analysis to their full 1975–1999 sample period, they continued to find substantial numbers of nominal wage cuts and a relative scarcity of nominal wage freezes.

Motivated by the onset and aftermath of the global financial crisis, our 2016 paper with Donggyun Shin replicated Nickell and Quintini's (2003) analysis and provided an update through the Great Recession to the year 2012. As summarized here in the second row of Table 1, our measured percentages of job stayers with

Table 1

Percentages of Job Stayers Receiving Year-to-Year Nominal Wage Cuts and Freezes

Study	*Data source*	*Wage measure*	*Percentage receiving wage cuts*	*Percentage receiving wage freezes*
Smith (2000)	British Household Panel Study, 1991–1996	Usual weekly pay from recent pay slip	17.8	5.6
Elsby, Shin, and Solon (2016)	British New Earnings Survey, 1975–2012	Earnings/hours excluding overtime for reference week in April	4.9[a]–23.5	0.4[a]–9.1
Jardim, Solon, and Vigdor (2019)	Washington State unemployment insurance records, 2005–2015	Quarterly earnings/hours	20.4–33.1	2.5–7.7
Bauer, Bonin, Goette, and Sunde (2007)	West German IABS-R[b] from social security records, 1975–1976, 1980–1981, . . . , 2000–2001	Annual earnings/work days for full-time workers employed on July 1	9.4–24.9	3.9–11.2
Evidence prepared for this survey by Andreas Steinhauer and Josef Zweimuller	Austrian Social Security Database, 2002–2012	Annual earnings/work days for full-time workers employed on March 15	13.0–18.6	0.1–1.5
Devicienti, Maida, and Sestito (2007)	Worker History Italian Panel from social security records, 1988–1989 and 1998–1999	Annual earnings/work days for full-time workers	7.7 and 18.3	4.0 and 8.5
Evidence prepared for OECD (2014) by Marcel Jansen, Sergi Jiménez, and José Ignacio García Pérez	Spanish Muestra Continua de Vidas Laborales from social security records, 2007–2010	Monthly earnings for full-time full-month workers	18.0–31.0	1.8–8.4
Castellanos, García-Verdú, and Kaplan (2004)	Mexican Social Security Institute records, 1985–2001	Daily comprehensive[c] wage on last day of quarter	0.2[a]–10.7	3.9[a]–16.5[d]
Doris, O'Neill, and Sweetman (2015)	Irish EU Survey of Income and Living Conditions, 2006–2011	Earnings/hours from recent pay slip for full-time full-year workers[e]	24.5–50.1	3.3–14.2
Park and Shin (2017)	South Korean Survey of Labor Conditions by Type of Employment, 2008–2013	Monthly earnings/hours excluding overtime and incentive pay in June	25.3–56.0	0.0–0.2

Continued on next page

nominal wage cuts ranged from a low of 4.9 percent in the period 1979–1980 (when inflation was around 20 percent) to a high of 23.5 percent in the wake of the Great Recession in both 2009–2010 and 2011–2012. Strikingly, the latter is by no means an aberration: over the last 20 years of the sample period, when the inflation rate in Britain hovered around 3 percent, the percentage of job stayers receiving nominal wage cuts was regularly close to 20 percent. Mirroring this impression of downward

Table 1 (Continued)

Percentages of Job Stayers Receiving Year-to-Year Nominal Wage Cuts and Freezes

Study	*Data source*	*Wage measure*	*Percentage receiving wage cuts*	*Percentage receiving wage freezes*
Carneiro, Portugal, and Varejão (2014)	Portuguese Quadros de Pessoal, 1986–1989, 1991–2000, and 2002–2016	Monthly base wage/ normal monthly hours for full-time workers in reference month[f]	2.2–6.3	3.2–76.0
Ekberg (2004)	Employer surveys by Confederation of Swedish Enterprise, 1970–1990 and 1995–1999	White-collar: Comprehensive[g] earnings/hours in reference month	White-collar: 0.1[a]–10.0	White-collar: 0.2[a]–6.0
		Blue-collar: Hourly base wage in second quarter	Blue-collar: 0.3[a]–3.9	Blue-collar: 0.0[a]–0.3
Vainiomäki (forthcoming)	Statistics Finland data based mostly on employer surveys by employer associations, 1995–2013	Earnings/hours excluding overtime in September, October, or fourth quarter	11.1–22.9	0.3–17.1

Note: Job stayers are defined as workers staying with the same employer; the British, Irish, Korean, Swedish, and Finnish studies also require that the workers stay in the same job within the firm.

[a] These data points correspond to periods of high inflation. They relate to 1979–1980 for Great Britain, when the inflation rate reached 20 percent; a period of hyperinflation in Mexico in the 1980s; and a period from the mid-1970s to the early 1980s in Sweden when the inflation rate regularly reached double digits.

[b] The IABS-R is part of the German Institute for Employment Research Employment Samples (IABS). It is a 2 percent random sample drawn from social security records.

[c] The Mexican wage measure "is a comprehensive measure of wages plus benefits, including payments made in cash, bonuses, premiums, room and board, commissions, benefits in kind and any other amount paid or benefit received."

[d] This excludes three outliers in the periods 1991:4–1992:4, 1996:4–1997:4, and 1998:4–1999:4, when increases in nominal minimum wages were not synchronized with the reporting dates. In each of these cases, the incidence of wage freezes exceeded 30 percent, at the expense of similar declines in the incidence of wage increases.

[e] The results from pay slips on earnings per hour are not reported in Doris, O'Neill, and Sweetman (2015), but they were kindly provided to us by Aedin Doris.

[f] Additional results not reported in Carneiro, Portugal, and Varejão (2014) were kindly provided to us by Pedro Portugal.

[g] The wage measure we cite for Swedish white-collar workers includes overtime, bonuses, and fringe benefits. Our reported percentage receiving wage cuts is a weighted average of the percentages Ekberg (2004) reports for white-collar workers who do and do not receive such supplementary payments.

flexibility, the incidence of zero nominal wage change was much smaller, varying from a low of 0.4 percent in the high-inflation period of 1979–1980 to a high of 9.1 percent in 2011–2012 and remaining below 3 percent in most years of the sample.

Like earlier researchers, we were intrigued by these findings, which motivated us to question whether similar studies might be feasible for other countries. As the remaining rows of Table 1 attest, it turns out that a body of such studies now exists, albeit one that has accumulated sporadically over a variety of journals spanning a range of fields of economics and that, in some cases, has become available only very recently.

United States

Although it is possible to access individual earnings data from some administrative sources in the United States, until recently it seemed that none contained the data on individual hours required to permit an analysis of hourly wages. However, thanks to the research of Kurmann, McEntarfer, and Spletzer (2016), considerable progress has been made on this seeming impasse. Their starting point was that US employers are obliged to report payroll data to state unemployment insurance agencies to enable determination of their employees' benefit entitlements in the event that the employees become unemployed and file an unemployment insurance claim. In most states, this requires employers to report the quarterly earnings of their employees. The key discovery by Kurmann, McEntarfer, and Spletzer was that a few states—Minnesota, Rhode Island, and Washington—also require employers to report their employees' quarterly hours of work. Among these, the case of Washington is especially useful because entitlement to unemployment insurance benefits in that state depends on hours as well as earnings, so the reports of both variables are thought to be especially accurate. Moreover, because these data are a near-complete census of employees in the state, they allow a researcher to track over time the wages of employees who remain with the same employer.

Two research teams—Kurmann and McEntarfer (2018) and Jardim, Solon, and Vigdor (2019)—have used the Washington data to study job stayers' year-to-year changes in quarterly average hourly earnings, and both have obtained results similar to those in the British studies. The third row of Table 1 summarizes the results from Jardim, Solon, and Vigdor, which are for the period 2005–2015. This period includes years before, during, and after the Great Recession, so although inflation was moderate throughout the period, business cycle conditions were wildly variable. Even during the expansion periods, the percentage receiving nominal wage cuts was more than 20 percent, with a minimum of 20.4 percent between the first quarters of 2006 and 2007. The percentage rose even higher during the Great Recession, with a high of 33.1 percent between the fourth quarters of 2008 and 2009. Mirroring this, the percentage receiving no nominal wage change typically remained below 4 percent, varying from a low of 2.5 percent between the fourth quarters of 2006 and 2007 to a maximum of just 7.7 percent at the height of the recession between the second quarters of 2009 and 2010. We are struck by the extent to which these results echo the British ones summarized above.

A contrast with the British studies using the New Earnings Survey, however, is that those studies were able to adopt a wage measure that explicitly excludes overtime pay and hours. Because overtime cannot be separated out in the Washington data, it is possible that some of the wage cuts measured for Washington could reflect reductions in overtime. As we noted above, these arguably should not be interpreted economically as wage reductions. Jardim, Solon, and Vigdor (2019) therefore redid their analysis for a subsample of workers who appeared to work 40 hours a week every week in each quarter. Even in this subsample, the frequency of nominal wage cuts was striking, ranging from a low of 14.5 percent between the third quarters of 2006 and 2007 to a high of 31.8 percent between the fourth quarters of 2008 and 2009.[4]

Evidence from Other Countries

Payroll records or pay slips have been used to study job stayers' nominal wage changes in many other countries, as shown in the remainder of Table 1. An Irish study included evidence similar to Smith's (2000) pay slip–based evidence for Great Britain. In Portugal and South Korea, the data were generated by government surveys of employers. In Sweden and Finland, the employer surveys were conducted by employer associations. As Table 1 documents, all of these studies allow an analysis of hourly wages similar to those we have summarized above for Great Britain and the United States.

In the studies for West Germany, Austria, Italy, Spain, and Mexico, the data are taken from employers' reports to their countries' social security systems. Since social security provisions typically do not require information on hours worked, most of these studies instead have focused on measurement of a daily wage. For West Germany, Austria, and Italy, this is computed as the ratio of annual earnings to days worked at a given employer. For Mexico, the daily wage is that measured on the last day of each quarter. Similarly, in Spain, the wage measure is based on monthly earnings for individuals who worked for the entire month. To allay concerns that changes in measured daily wages reflect changes in hours worked per day, all but one of these studies (the Mexican case) additionally focus on individuals recorded as working full time in the administrative data.

Not surprisingly, the patterns vary considerably across countries. We think it is a fair summary to say that, outside of conditions of very high price inflation, most of the countries continue to show substantial minorities of job stayers receiving nominal wage cuts and much smaller minorities experiencing zero nominal wage change.

According to the Italian study by Devicienti, Maida, and Sestito (2007), for example, in the period 1988–1989, when inflation was a relatively high 6.5 percent, the percentage receiving nominal wage cuts was "only" 7.7 percent. In the period 1998–1999, when inflation was under 2 percent, the percentage receiving wage cuts was 18.3 percent. Qualitatively similar results are reported for West Germany by Bauer, Bonin, Goette, and Sunde (2007) and for Spain by the OECD (2014), except that the percentage receiving wage cuts ran somewhat higher, peaking at 24.9 percent in 1995–1996 for West Germany and at 31.0 percent in 2009–2010 in

[4] A preliminary manuscript by Grigsby, Hurst, and Yildirmaz (2018) that uses US data from the ADP payroll processing company finds that base pay reductions are rare in expansion years, but that reductions in overall earnings per hour are strikingly common, even with overtime excluded. This finding regarding the role of compensation other than base pay (such as bonuses) in nominal wage adjustment echoes a similar finding in the literature on cyclicality in *real* wages (see Shin and Solon 2007 and the references therein).

the aftermath of the especially severe Great Recession in Spain. For all three countries, the percentage of job stayers recorded with no wage change never rose much above 10 percent.

The Austrian evidence, kindly prepared for this survey by Andreas Steinhauer and Josef Zweimuller, again points to a considerable prevalence of nominal wage cuts. Over a 2002–2012 sample period that rarely saw inflation rise above 3 percent, the percentage receiving nominal wage cuts ranged from 13.0 to 18.6 percent. Strikingly, nominal wage freezes were exceedingly rare in the Austrian data, affecting less than 2 percent of job stayers. Vainiomäki's (forthcoming) results for Finland are fairly similar. The percentage receiving nominal wage cuts was always more than 11 percent and was usually more than 15 percent. In all but two of the sample period's 18 years, the percentage with wage freezes was 5 percent or less.

Inflation plays a particularly important role in the Mexican results reported by Castellanos, García-Verdú, and Kaplan (2004). In the early part of their 1985–2001 sample period, when annual inflation soared (reaching almost 160 percent!), nominal wage cuts were extremely rare. At the end, when inflation was just starting to moderate to single digits, the percentage receiving wage cuts had risen to 10.7 percent. At the same time, aside from a few periods in which rises in the nominal minimum wage were delayed, no more than 16.5 percent of job stayers experienced no change in their nominal wage.

The outliers in Table 1 are especially instructive. At one extreme are the results reported by Doris, O'Neill, and Sweetman (2015) for Ireland, where the Great Recession hit especially hard and involved a price deflation. In the period 2009–2010, the percentage of job stayers receiving nominal wage cuts reached a striking 50.1 percent. Even in the depths of the crisis in Ireland, the incidence of nominal wage freezes rose no higher than 14.2 percent.

In their results for South Korea, Park and Shin (2017) report a similarly extreme frequency of wage cuts, which affected as much as 56.0 percent of job stayers in 2008–2009, when both output growth and inflation were close to zero. An equally striking aspect of the South Korean data, however, is that the percentage of job stayers experiencing zero change in their nominal wage was negligible. The data for South Korea thus exhibit none of the empirical hallmarks of downward nominal wage rigidity, in precisely the macroeconomic context in which one might expect to find them.[5]

At the other extreme is Portugal, where Carneiro, Portugal, and Varejão (2014) report that nominal wage cuts were "virtually non-existent" throughout the 1987–2009 period, affecting no more than 6 percent of job stayers. This makes sense because Portugal has a national law that explicitly prohibits such cuts. Consistent with this, the incidence of nominal wage freezes in Portugal rose to unparalleled levels during the Great Recession, when zero change in hourly pay was recorded for up to 76.0 percent of job stayers.

[5] A newer study by Park and Shin (forthcoming) extends their evidence back to 1986.

At first blush, the situation seems somewhat similar in Sweden. For blue-collar workers, Ekberg (2004, chap. 1) reports that between 0.3 and 3.9 percent received hourly base wage cuts. He explains that, "given the framework of the terms of employment, it is impossible for the employers to cut wages unilaterally. Hence, a wage cut can only be achieved under mutual consent," and even then it cannot violate applicable collective bargaining agreements. In stark contrast to the Portuguese case, however, almost none of these Swedish job stayers experienced a nominal pay freeze. Moreover, although Ekberg reports very low percentages of white-collar workers with wage cuts at the beginning of his sample period (when inflation was in double digits and very few white-collar workers received any supplementary pay), by the end inflation was much lower, a majority of white-collar workers received some supplementary pay, and the percentage receiving pay cuts rose as high as 10.0 percent.

Figure 1 supplements Table 1 by providing a visual representation of the frequency of nominal wage cuts as a function of inflation. For the sake of a readable scale, the figure excludes the Mexican observations in which the inflation rate exceeds 20 percent (sometimes by a lot!) and the associated frequency of nominal wage cuts is negligible. Like Table 1, Figure 1 indicates that, outside of periods of particularly high inflation, most countries exhibit surprisingly high frequencies of nominal wage cuts. In addition, the figure reveals a general tendency for the frequency of wage cuts to rise as inflation falls. The glaring exception is Portugal, where a national prohibition of nominal wage cuts makes it the canonical example of Keynes's premise that nominal wages cannot be cut. As discussed above, while nominal wage cuts appear to be rare in Sweden as well, there is little evidence there for an associated buildup of wage freezes. Otherwise, the evidence accumulated from payroll records and pay slips suggests that nominal wage cuts occur more commonly than most of us had thought.

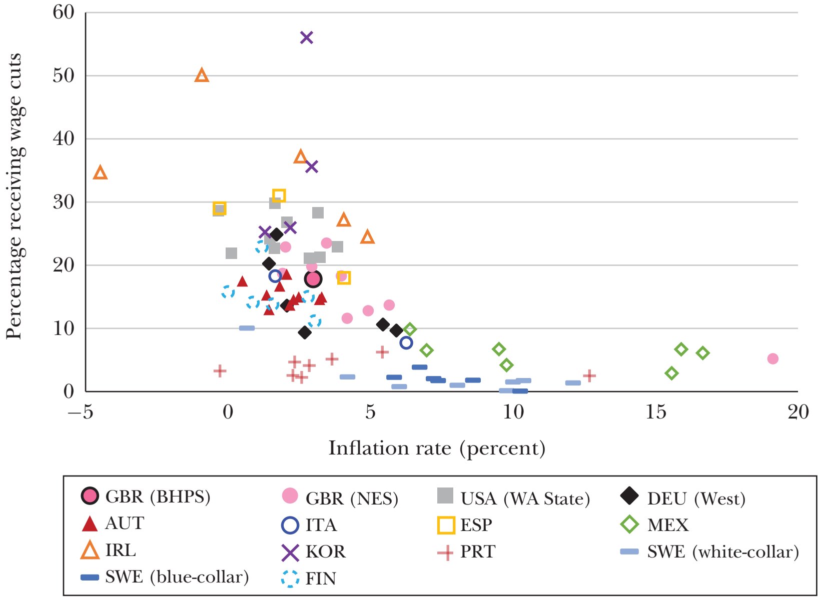
Some Nuances

Having found that nominal wage decreases occur with surprising frequency, we can inquire further about how they are distributed throughout the labor market. Recent findings suggest that the overall flexibility we report is pervasive, in two senses.

First, Elsby, Shin, and Solon (2016) point out that the nominal wage cuts observed in the British New Earnings Survey "are remarkably pervasive across sub-groups of workers/jobs. For example, in 2011–2012, when the overall proportion of job stayers experiencing cuts was 23.5%, the proportions were 22% in the private sector and 26% in the public sector; 27% for union workers and 22% for nonunion workers; at least 20% for every single-digit occupation; and 32% for workers who received incentive pay in either 2011 or 2012 and 22% for workers who did not." The study of Washington State data by Jardim, Solon, and Vigdor (2019) also presents some disaggregated analyses, and it similarly finds that the common occurrence of nominal wage cuts is pervasive across both industries and firm sizes. Even in the utilities industry—the industry that tends to show the fewest

Figure 1

Percentages of Job Stayers Receiving Year-to-Year Nominal Wage Cuts as a Function of Inflation



Notes: This figure provides a visual representation of the frequency of nominal wage cuts as a function of inflation based on the literature survey summarized in Table 1. Inflation rates corresponding to the NES data for Great Britain are from Elsby, Shin, and Solon (2016). Inflation rates for all other studies are from OECD data (https://data.oecd.org/price/inflation-cpi.htm). For studies with annual data, corresponding annual inflation rates are used. For studies with quarterly data, corresponding quarterly inflation rates are used and then simple annual averages are taken. For studies with many years of data, the figure plots a selected sample of years, chosen to include both the minimum and maximum percentage of wage cuts reported in Table 1 and otherwise evenly sampled across the available years. Finally, the figure focuses on periods for each study in which the inflation rate was no greater than 20 percent. Country abbreviations are OECD country codes. Other abbreviations: BHPS, British Household Panel Study; NES, New Earnings Survey.

nominal wage cuts—the percentage receiving cuts was almost always 15 percent or greater.[6]

Second, recent studies with access to rich employer-employee matched data have begun to investigate whether firms cutting wages do so for nearly all their workers or target the cuts on selected subgroups. For example, if 20 percent of all the job stayers in a particular period show wage cuts, this could happen because 20 percent of the stayers in every firm receive wage cuts or because the cuts occur

[6] Another type of heterogeneity that future research could explore is with respect to whether economic shocks are general or idiosyncratic to the firm. The recent study by Juhn, McCue, Monti, and Pierce (2018) concludes that "the transmission of firm-level shocks to earnings of stayers is minimal in the US labor market."

universally in firms that employ 20 percent of stayers but not at all in other firms. Where between these extremes does the reality lie? To explore this question with the Washington State data, Jardim, Solon, and Vigdor (2019) created for each job stayer receiving a wage cut the following variable: the percentage of that worker's job-staying coworkers that also received a wage cut in the same period. In every period studied, it turned out that the majority of job stayers receiving nominal wage cuts worked for firms that cut the wages of between 10 and 50 percent of their job stayers. Jardim, Solon, and Vigdor also noted a tendency for these selective wage cuts to be more concentrated in the upper half of within-firm wage distributions. Park and Shin (2017) have reported similar findings for South Korea, noting that the prevalence of nominal wage cuts summarized in Table 1 stems from "a majority of employers cutting a fraction of their workers' wages fairly routinely."

We regard these details as promising points of departure for further research. They suggest that nominal wage cuts are not only surprisingly common but also broadly distributed across sectors and firms.

Summary and Discussion

For more than 80 years, many (though far from all) influential macroeconomic analyses of the labor market have been premised on the assumption that nominal wages cannot be cut. Some classic studies that used longitudinal household surveys to track job stayers from year to year measured a high incidence of wage cuts, but this evidence reasonably was discounted on the grounds that the measurement of frequent wage cuts could be an artifact of survey response error.

The main point of the present paper has been to synthesize a more recent international collection of studies that have sought out more accurate wage data from employers' payroll records and pay slips. Outside of circumstances where nominal wage cuts have been legally prohibited or rendered irrelevant by very high price inflation, most of this evidence has continued to show that nominal wage cuts occur more frequently than has commonly been supposed.

Most of us are surprised by this finding, not only because of the persistent influence of Keynes's (1936) contrary assumption in *The General Theory* but also because introspection, casual empiricism, and Bewley's (1999) interviews tell us that workers really do dislike nominal wage cuts and employers are therefore reluctant to impose them. But is this obvious aversion to wage cuts so extreme as to bind even when inefficient layoffs into unemployment are the alternative? The accumulated international evidence showing that nominal wage cuts occur frequently should inspire reconsideration of the commonly invoked assumption that nominal wages *cannot* be cut even when efficiency of allocation decisions is at stake.

Of course, because the evidence reviewed here is based on longitudinal tracking of job stayers, it pertains directly only to wage rigidity for incumbent workers. As discussed above, a related question is how flexible wages are for the hiring of new workers. Some recent models have assumed that wage rigidity for incumbents spills over into wage rigidity for new hires. In that light, the evidence reported here is indirectly pertinent for hiring wages. If nominal wage cuts are feasible for incumbent workers, why would they not be for new workers?

The development of theoretically coherent and empirically relevant accounts of what happens in the labor market over the business cycle remains a crucial mission for economic research. We hope to support that effort by providing a more accurate picture of the frequency and nature of nominal wage cuts.

■ *We gratefully acknowledge financial support from the UK Economic and Social Research Council, award reference ES/L990633/1. We thank the editors for guidance and Francesco Devicienti, Aedin Doris, José Ignacio García Pérez, David Kaplan, Pedro Portugal, Uwe Sunde, and Jari Vainiomäki for generously providing additional details on the results from their studies. We also thank Andreas Steinhauer and Josef Zweimuller for kindly providing the results for Austria.*

References

Akerlof, George A., William T. Dickens, and George L. Perry. 1996. "The Macroeconomics of Low Inflation." *Brookings Papers on Economic Activity*, no. 1, pp. 1–76.

Altonji, Joseph G., and Paul J. Devereux. 1999. "The Extent and Consequences of Downward Wage Rigidity." NBER Working Paper 7236.

Barro, Robert J. 1977. "Long-Term Contracting, Sticky Prices, and Monetary Policy." *Journal of Monetary Economics* 3(3): 305–16.

Bauer, Thomas, Holger Bonin, Lorenz Goette, and Uwe Sunde. 2007. "Real and Nominal Wage Rigidities and the Rate of Inflation: Evidence from West German Micro Data." *Economic Journal* 117(524): F508–29.

Becker, Gary S. 1962. "Investment in Human Capital: A Theoretical Analysis." *Journal of Political Economy* 70(5): 9–49.

Benigno, Pierpaolo, and Luca Antonio Ricci. 2011. "The Inflation-Output Trade-Off with Downward Wage Rigidities." *American Economic Review* 101(4): 1436–66.

Bewley, Truman F. 1999. *Why Wages Don't Fall during a Recession*. Cambridge, MA: Harvard University Press.

Card, David, and Dean Hyslop. 1996. "Does Inflation 'Grease the Wheels of the Labor Market?'" NBER Working Paper 5538.

Carneiro, Anabela, Pedro Portugal, and José Varejão. 2014. "Catastrophic Job Destruction during the Portuguese Economic Crisis." *Journal of Macroeconomics* 39(B): 444–57.

Castellanos, Sara G., Rodrigo García-Verdú, and David S. Kaplan. 2004. "Nominal Wage Rigidities in Mexico: Evidence from Social Security Records." *Journal of Development Economics* 75(2): 507–33.

Daly, Mary C., and Bart Hobijn. 2014. "Downward Nominal Wage Rigidities Bend the Phillips Curve." *Journal of Money, Credit and Banking* 46(S2): 51–93.

Devicienti, Francesco, Agata Maida, and Paolo Sestito. 2007. "Downward Wage Rigidity in Italy: Micro-Based Measures and Implications." *Economic Journal* 117(524): F530–52.

Dickens, William T., Lorenz Goette, Erica L. Groshen, Steinar Holden, Julian Messina, Mark E. Schweitzer, Jarkko Turunen, and Melanie E. Ward. 2007. "How Wages Change: Micro Evidence from the International Wage Flexibility Project." *Journal of Economic Perspectives* 21(2): 195–214.

Doris, Aedin, Donal O'Neill, and Olive Sweetman. 2015. "Wage Flexibility and the Great

Recession: The Response of the Irish Labour Market." *IZA Journal of European Labor Studies* 4(1): 18.

Dupraz, Stéphane, Emi Nakamura, and Jón Steinsson. 2018. "A Plucking Model of Business Cycles." December 20. https://eml.berkeley.edu/~enakamura/papers/plucking.pdf.

Ekberg, John. 2004. "Essays in Empirical Labor Economics." PhD Dissertation, Department of Economics, Stockholm University.

Elsby, Michael W. L., Donggyun Shin, and Gary Solon. 2016. "Wage Adjustment in the Great Recession and Other Downturns: Evidence from the United States and Great Britain." *Journal of Labor Economics* 34(S1): S249–91.

Gertler, Mark, and Antonella Trigari. 2009. "Unemployment Fluctuations with Staggered Nash Wage Bargaining." *Journal of Political Economy* 117(1): 38–86.

Grigsby, John, Erik Hurst, and Ahu Yildirmaz. 2018. "Aggregate Nominal Wage Adjustments: New Evidence from Administrative Payroll Data." Unpublished paper.

Hall, Robert E. 2005. "Employment Fluctuations with Equilibrium Wage Stickiness." *American Economic Review* 95(1): 50–65.

Jardim, Ekaterina S., Gary Solon, and Jacob L. Vigdor. 2019. "How Prevalent Is Downward Rigidity in Nominal Wages? Evidence from Payroll Records in Washington State." NBER Working Paper 25470.

Juhn, Chinhui, Kristin McCue, Holly Monti, and Brooks Pierce. 2018. "Firm Performance and the Volatility of Worker Earnings." *Journal of Labor Economics* 36(S1): S99–131.

Kahn, Shulamit. 1997. "Evidence of Nominal Wage Stickiness from Microdata." *American Economic Review* 87(5): 993–1008.

Keynes, John Maynard. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Kurmann, André, and Erika McEntarfer. 2018. "Downward Wage Rigidity in the United States: New Evidence from Administrative Data." Unpublished paper. http://www.andrekurmann.com/files/wp_files/KM_December2018_final.pdf.

Kurmann, André, Erika McEntarfer, and James Spletzer. 2016. "Downward Wage Rigidity in the U.S.: New Evidence from Worker-Firm Linked Data." Unpublished paper. http://www.andrekurmann.com/files/wp_files/KMS_20160215.pdf.

Lebow, David E., Raven E. Saks, and Beth Anne Wilson. 2003. "Downward Nominal Wage Rigidity: Evidence from the Employment Cost Index." *Advances in Macroeconomics* 3(1).

Malcomson, James M. 1997. "Contracts, Hold-Up, and Labor Markets." *Journal of Economic Literature* 35(4): 1916–57.

Martins, Pedro S., Gary Solon, and Jonathan P. Thomas. 2012. "Measuring What Employers Do about Entry Wages over the Business Cycle: A New Approach." *American Economic Journal: Macroeconomics* 4(4): 36–55.

McLaughlin, Kenneth J. 1994. "Rigid Wages?" *Journal of Monetary Economics* 34(3): 383–414.

Nickell, Stephen, and Glenda Quintini. 2003. "Nominal Wage Rigidity and the Rate of Inflation." *Economic Journal* 113(490): 762–81.

OECD. 2014. "Sharing the Pain Equally? Wage Adjustments during the Crisis and Recovery." Chap. 2 in *OECD Employment Outlook 2014*. Paris: OECD Publishing.

Park, Seonyoung, and Donggyun Shin. 2017. "The Extent and Nature of Downward Nominal Wage Flexibility: An Analysis of Longitudinal Worker/Establishment Data from Korea." *Labour Economics* 48(October): 67–86.

Park, Seonyoung, and Donggyun Shin. Forthcoming. "Inflation and Wage Rigidity/Flexibility in the Short Run." *Economic Inquiry*.

Schmitt-Grohé, Stephanie, and Martín Uribe. 2016. "Downward Nominal Wage Rigidity, Currency Pegs, and Involuntary Unemployment." *Journal of Political Economy* 124(5): 1466–514.

Shimer, Robert. 2004. "The Consequences of Rigid Wages in Search Models." *Journal of the European Economic Association* 2(2/3): 469–79.

Shin, Donggyun, and Gary Solon. 2007. "New Evidence on Real Wage Cyclicality within Employer-Employee Matches." *Scottish Journal of Political Economy* 54(5): 648–60.

Smith, Jennifer C. 2000. "Nominal Wage Rigidity in the United Kingdom." *Economic Journal* 110(462): 176–95.

Snell, Andy, and Jonathan P. Thomas. 2010. "Labor Contracts, Equal Treatment, and Wage-Unemployment Dynamics." *American Economic Journal: Macroeconomics* 2(3): 98–127.

Tobin, James. 1972. "Inflation and Unemployment." *American Economic Review* 62(1/2): 1–18.

Vainiomäki, Jari. Forthcoming. "The Development of Wage Dispersion and Wage Rigidity in Finland." *Finnish Economic Papers* 29(1).

Should We Tax Sugar-Sweetened Beverages? An Overview of Theory and Evidence

Hunt Allcott, Benjamin B. Lockwood,
and Dmitry Taubinsky

"Sin taxes" are imposed to discourage individual behaviors, such as smoking or drinking alcohol, that are thought to harm the individual and possibly others in society. This article provides an economic framework for evaluating an increasingly popular class of sin taxes: those on sugar-sweetened beverages. As of mid-2019, seven US cities and thirty-nine countries around the world have implemented sugar-sweetened beverage taxes, mostly in the past few years (Global Food Research Program 2019).

Proponents of these taxes point to a range of policy goals, including improving public health and raising revenues that can be used to reduce budget deficits or to fund social programs. Opponents often express concerns about paternalistic government intervention in individual decisions and point out that sugar-sweetened beverages are consumed most heavily by the poor, which could make taxes regressive. How do economists evaluate these arguments? Should we tax sugar-sweetened beverages? If so, how high should the tax be?

■ *Hunt Allcott is Associate Professor of Economics, New York University, New York, New York, and Principal Researcher, Microsoft Research, Cambridge, Massachusetts. Benjamin B. Lockwood is Assistant Professor of Business Economics and Public Policy, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania. Dmitry Taubinsky is Assistant Professor of Economics, University of California, Berkeley, California. Allcott is a Research Associate and Lockwood and Taubinsky are Research Fellows, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are hunt.allcott@nyu.edu, ben.lockwood@wharton.upenn.edu, and dmitry.taubinsky@berkeley.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.202 doi=10.1257/jep.33.3.202

Table 1
Sugar-Sweetened Beverage Taxes in the United States

Location	*Date enacted*	*Tax rate (¢ per ounce)*	*Includes diet drinks?*
Albany, CA	November 2016	1	No
Berkeley, CA	November 2014	1	No
Boulder, CO	November 2016	2	No
Oakland, CA	November 2016	1	No
Philadelphia, PA	June 2016	1.5	Yes
San Francisco, CA	November 2016	1	No
Seattle, WA	June 2017	1.75	No
Cook County, IL	November 2016 (repealed October 2017)	1	Yes

Source: Data obtained through the authors' research via municipal and county websites and Ballotpedia.

In the first part of the article, we provide background on sugar-sweetened beverage consumption patterns and the resulting health harms. This section helps to explain why sugary drinks have come to be seen as a "sin" worthy of taxation. In the second part of the article, we draw on our recent work (Allcott, Lockwood, and Taubinsky 2019) to present the economic principles that determine the optimal level of taxes on sugar-sweetened beverages. We discuss how the price elasticity of demand, externalities, "internalities," distributional concerns, and the incidence on producers all shape the optimal tax on sugar-sweetened beverages. In the third part of the article, we summarize the growing empirical literature that estimates these key parameters.

We end with seven concrete suggestions for policymakers. First, focus on counteracting externalities and internalities, not on minimizing sugary drink consumption. Second, target policies to reduce consumption among people generating the largest externalities and internalities. Third, tax grams of sugar, not ounces of liquid. Fourth, tax diet drinks and fruit juice if and only if they also cause uninternalized health harms. Fifth, when judging regressivity, consider internality benefits, not just who pays the taxes. Sixth, if possible, implement taxes statewide. Finally, the benefits of sugar-sweetened beverage taxes probably exceed their costs.

Background: Taxes, Consumption, and Health Harms

Existing Sugar-Sweetened Beverage Taxes

Table 1 presents the seven current city-level sugar-sweetened beverage taxes in the United States, all of which have been enacted since 2014. Cook County, Illinois, which contains the city of Chicago, passed a tax and then repealed it a year later. The modal tax rate is 1 cent per ounce, although Boulder, Philadelphia, and Seattle have higher rates. In addition to these explicit taxes, 23 states plus the District of Columbia exempt or partially exempt groceries from sales taxes but do not define sugar-sweetened beverages as "groceries," thereby taxing these drinks at a higher

Table 2
Sugar-Sweetened Beverage Taxes around the World

Europe	*Western Pacific*	*Africa, Eastern Mediterranean, and Southeast Asia*	*Americas*
Estonia (2018)	Philippines (2018)	Morocco (2019)	Colombia (2019)
Ireland (2018)	Brunei (2017)	South Africa (2018)	Bermuda (2018)
United Kingdom (2018)	Vanuatu (2015)	Bahrain (2017)	Peru (2018)
Portugal (2017)	Kiribati (2014)	India (2017)	Barbados (2015)
Belgium (2016)	Cook Islands (2013)	Maldives (2017)	Dominica (2015)
France (2012)	Tonga (2013)	Sri Lanka (2017)	Chile (2014)
Hungary (2011)	Fiji (2007)	Saudi Arabia (2017)	Mexico (2014)
Latvia (2004)	Nauru (2007)	Thailand (2017)	
Norway (1981)	Palau (2003)	United Arab Emirates (2017)	
Finland (1940)	French Polynesia (2002)	St. Helena (2014)	
	Samoa (1984)	Mauritius (2013)	

Source: Based on data from the Global Food Research Program (2019).
Notes: The table lists countries with taxes on sugar-sweetened beverages, grouped by region; year of implementation is given in parentheses.

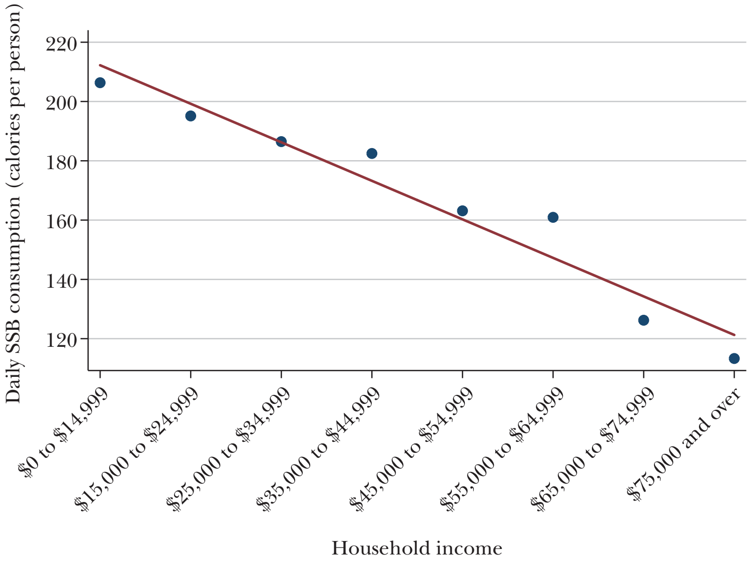
rate (Loughead 2018). Table 2 lists the 39 other countries around the world with sugar-sweetened beverage taxes.

A sugar-sweetened beverage is any drink with caloric sweeteners, including carbonated soft drinks, sports drinks, energy drinks, fruit drinks, chocolate (or otherwise sweetened) milk, and sweetened coffee and tea, but not including 100 percent fruit juice or "diet" drink alternatives with noncaloric sweeteners. The beverage categories included in sugary drink taxes depend on both political calculations and judgment calls by public health experts. All the city-level taxes in Table 1 cover all sugar-sweetened beverages except for sweetened milk products, and they do not include 100 percent fruit juice, on the grounds that the vitamins and nutrients such as calcium in these drinks provide some additional nutritional value. The Philadelphia tax and the repealed Cook County tax additionally include diet drinks. As we discuss below, it is not clear that these coverage decisions are socially optimal.

Sugar-Sweetened Beverage Consumption

Americans consume a remarkable amount of calories from sugar-sweetened beverages. A typical 12-ounce soft drink might contain 35 to 40 grams of sugar and about 140 calories, representing about 7 percent of a benchmark diet of 2,000 calories per day. Using data from the National Health and Nutrition Examination Survey (NHANES) for the period 2009–2016, we calculate that the average American adult consumes 154 calories per day from sugar-sweetened beverages, which represents 6.9 percent of actual total calorie intake. Almost all of these calories are from added sugars. As a benchmark, the US Dietary Guidelines recommend limiting added sugars from all food and drinks to no more than 10 percent of total calorie intake, or around 200 calories per day, while the World Health Organization is even more conservative. In the NHANES data, sugar-sweetened beverages account for 23 percent of the average American adult's total sugar consumption.

Figure 1
Sugar-Sweetened Beverage Consumption by Income



Source: Authors, using data from the National Health and Nutrition Examination Survey.
Notes: Figure shows average daily sugar-sweetened beverage (SSB) consumption by household income for the period 2009–2016. Consumption varies substantially by income; on average, people with higher household income consume fewer calories from sugar-sweetened beverages, which raises concerns that taxes on such drinks could be regressive.

Sugary drinks are broadly popular: about 50 percent of American adults consume at least one sugar-sweetened beverage on any given day. However, Figure 1 shows that consumption varies substantially by income. People with household income below $25,000 per year consume 200 calories per day of sugar-sweetened beverages, while people with household income above $75,000 per year consume only 117 calories per day. This generates the concern that taxes on sugar-sweetened beverages could be regressive. There is also substantial within-group variation: in a large nationwide survey carried out by Nielsen for Allcott, Lockwood, and Taubinsky (2019), the 90th percentile of individual consumption is 2.7 times as large as the mean, and 6.5 times as large as the median.

Perhaps due to rising public awareness of the health effects of sugar-sweetened beverages, consumption is falling over time in the United States and many other Western countries. In the National Health and Nutrition Examination Survey data, the average American consumed 205 calories per day from sugar-sweetened beverages in 2003–2004, compared with 154 calories in 2009–2016. Popkin and Hawkes (2016) find that sugar-sweetened beverage calorie consumption per capita declined from 2009 to 2014 in North America, Australasia, and Western Europe but increased in the rest of the world. They also report that North Americans consume 3 to 4 times more calories from sugar-sweetened beverages than the world average.

Health Harms from Sugar-Sweetened Beverage Consumption

Sugar-sweetened beverage consumption harms health through three main channels: weight gain, type 2 diabetes, and cardiovascular disease. (We do not discuss other health effects, such as tooth decay.) For these main channels, we briefly discuss evidence on the magnitude of the effects. Some of this evidence comes from nonrandomized epidemiological studies that correlate sugar-sweetened beverage consumption with health outcomes, while attempting to control for confounding variables. Although this is sometimes the only evidence available, correlation doesn't imply causation: additional unmeasured confounders such as eating patterns, exercise, and social conditions could mean that these conditional correlations are inaccurate measures of the causal effect of sugar-sweetened beverage consumption on health. Moreover, most quantitative studies report only an average effect, though the effects may be concentrated on particular groups or heavy users.

The first main health harm is weight gain. Some evidence suggests that sugary drinks cause more weight gain than equally sugary foods because calories are less satiating in liquid form (Pan and Hu 2011). A randomized experiment by Mourao et al. (2007) found that when people consume the same amount of calories from solid foods instead of liquids (say, jelly beans instead of soda, or cheese instead of milk), they eat less later in the day, resulting in significantly lower overall calorie intake. Other experiments have found that when people are provided with the same foods and either caloric or noncaloric beverages, they consume the same amount of calories from food regardless of the beverage provided and report no difference in feelings of satiety (DellaValle, Roe, and Rolls 2005; Flood, Roe, and Rolls 2006).

Both field experiments and nonexperimental analyses have estimated the weight gain effects. Randomized trials with children and adolescents find that substituting diet drinks for sugar-sweetened beverages for 12 to 18 months reduces weight by 2 to 4 pounds (de Ruyter et al. 2012; Ebbeling et al. 2012). In observational analysis of three adult cohort studies, Mozaffarian et al. (2011) find that one additional serving per day of sugar-sweetened beverages is conditionally associated with weight gain of 1 pound per four-year follow-up period, after controlling for a variety of biological and lifestyle factors.[1]

The second main health harm is type 2 diabetes. Sugar-sweetened beverages have high "glycemic loads," meaning that they contain large amounts of rapidly digestible sugars. Sugars are digested more quickly when they come from drinks than when they are eaten with food. When foods or drinks with high glycemic loads are digested, they prompt a quick release of glucose into the bloodstream and the secretion of a corresponding amount of insulin in response. Over time, these states of elevated blood glucose and insulin can cause insulin resistance, often a precursor

[1] For a review of additional randomized experiments on the effects of sugar-sweetened beverages on weight gain, see Mattes et al. (2011). For reviews of cohort studies, see Vartanian, Schwartz, and Brownell (2007) and Malik et al. (2013). Weight gain is thought to have an independent effect on diabetes and cardiovascular disease in addition to the mechanisms described below, and weight mediates the statistical relationships between sugar-sweetened beverage consumption and those conditions (for example, Schulze et al. 2004; Fung et al. 2009).

to diabetes (see description in Ludwig 2002; Raben et al. 2011). A meta-analysis of 17 cohort studies found that drinking one more serving of sugar-sweetened beverages per day was associated with a 13 percent higher risk of developing type 2 diabetes (Imamura et al. 2015; see also Malik et al. 2010).

The third main health harm is cardiovascular diseases, such as heart attack and narrowing of the arteries. Randomized trials show that diets high in sugar and other refined carbohydrates increase blood pressure and cholesterol; high blood pressure and high cholesterol are precursors to cardiovascular disease (Santos et al. 2012; Te Morenga et al. 2014). A meta-analysis of four studies found that consuming one additional serving of sugar-sweetened beverages per day is associated with a 17 percent higher risk of coronary heart disease (Xi et al. 2015).

This background helps explain why the public health community has focused on taxing sugary drinks instead of a broader sugar tax that includes sugar in foods: sugar consumed through drinks is more harmful.

Quantifying Health System Costs

By combining estimates of the price elasticity of demand for sugar-sweetened beverages, the effect of sugar-sweetened beverages on diabetes, cardiovascular disease, and obesity, and the costs of treating these diseases, it is possible to estimate the effects of a sugar-sweetened beverage tax on health-care costs. The necessary parameters are often estimated from correlation studies and are thus subject to the same important caveat that correlation does not imply causation. However, Wang et al. (2012) estimate that over 10 years, a 1 cent per ounce tax would save $17.1 billion in health-care costs. Using a separate model, Long et al. (2015) estimate the ten-year savings to be $23.6 billion.

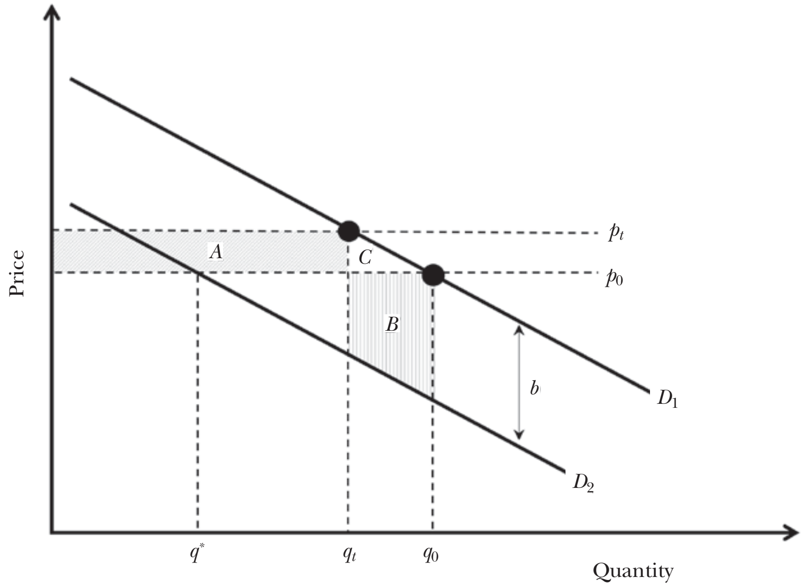
An Economic Framework for Evaluating Sugar-Sweetened Beverage Taxes

The economic logic behind a tax on sugar-sweetened beverages builds from the classic principles of externality-correcting taxes (Pigou 1920): if consuming a good harms others, then people will consume too much if the market is not regulated. Thus, a tax imposed on a good with negative externalities can raise welfare by reducing consumption toward the efficient level at which marginal social cost equals marginal social benefit.

Additionally, a growing body of research in behavioral economics indicates that people sometimes ignore harmful or beneficial effects to *themselves*—for example, because they are misinformed, or because they do not fully consider future health consequences due to “present focus.” These costs are sometimes called “internalities,” and we view their presence as a key distinction in the rationale for “sin taxes” on goods like cigarettes and alcohol.

It is important to emphasize that externalities and internalities are not the same as “health harms.” A consumer might rationally drink something (or take any

Figure 2
Effect of a Sugar-Sweetened Beverage Tax on Individual Consumption



Source: The authors; see Allcott, Lockwood, and Taubinsky (2019) for a formal treatment.
Notes: The figure illustrates the effect of a sugar-sweetened beverage tax on demand from a single consumer. D_1 plots the individual's demand curve for sugar-sweetened beverages at various prices. The vertical distance b represents the per unit externality or internality cost (see text for details).

other action) despite the health risks, because enjoyment of the drink outweighs the health harms. What matters for sin taxes is whether consumers' choices impose harms on others (externalities) or harms on themselves that they do not correctly internalize (internalities).

Although internality and externality costs operate somewhat similarly, there are important differences between the two, and we consider each in turn. Figure 2, which illustrates the effect of a sugar-sweetened beverage tax on demand from a single consumer, can be used to discuss both concepts. In Allcott, Lockwood, and Taubinsky (2019), we provide a formal treatment of the issues this section.

Welfare Effects Due to Externalities

Some sin goods generate direct consumption externalities—cigarettes create second-hand smoke, for example. In the context of sugar-sweetened beverages, probably the most important externalized cost is in the form of financial health-care costs, which are shared through public or private insurance. Strictly speaking, these are moral hazard costs, or "fiscal externalities" (in the case of public insurance), which arise due to preexisting information frictions in a second-best world. We will call all such externalized costs "externalities," however, to emphasize that they are borne by people other than the consumer of sugar-sweetened beverages.

In Figure 2, to illustrate the role of externalities, D_1 plots the individual's demand curve for sugar-sweetened beverages at various prices (or, equivalently, the consumer's marginal private benefit from sugary drinks at each quantity). The vertical distance b represents the per unit externality cost, so that D_2 plots the marginal social benefit from consumption, net of externalities, as a function of quantity consumed. (In practice, b may vary with the level of consumption; here we plot it as a constant marginal externality for simplicity.) A tax that raises the price from p_0 to p_t then has three distinct effects on welfare. (For simplicity here we assume the tax is fully passed through to consumers—we relax that assumption below.) The area $A = t \times q_t$ is transferred from the consumer to the government, in the form of tax revenue. The area $C = \Delta q \times t/2$ represents a further decrease in the consumer's welfare from foregone sugary drink consumption due to the tax. The area $B + C = \Delta q \times b$ represents an increase in welfare for those bearing the externality. In the context of sugar-sweetened beverages, a natural benchmark assumption is that the externality reduction accrues to the government's budget (in present value terms)—for example, due to reduced Medicare expenditures on treatments for conditions such as heart disease and diabetes. Therefore, the net effect of the tax is twofold: a transfer of $A + C$ from the consumer to the government, and a further increase in government funds of B.

The total welfare effects of an externality-based sugar-sweetened beverage tax depend on aggregating these components across individuals. Because the tax involves transfers between parties, something must be assumed about the social value of resources in the hands of the government relative to consumers, and across consumers of different types. A common assumption is that the marginal utility from consumption is decreasing with consumers' incomes—the same assumption that is often used to justify progressive income tax schedules. One way to capture such distributional implications is to assign "social marginal welfare weights" (as in Saez and Stantcheva 2016) to different households depending on their income (or possibly other attributes), so that a weight of, say, 1.5 on household x implies that society places the same value on \$1 in the hands of household x as on \$1.50 in the hands of the government. Then the transfer $A + C$ from the consumer to the government generates a net social gain if the weight on the consumer in question is less than 1, and a social loss otherwise.

Putting these pieces together, we aggregate these effects by summing the externality benefit B and the transfer $A + C$ across consumers, weighted appropriately. The area B scales with its width (proportional to the individual elasticity of demand for sugary drinks) multiplied by its height (the externalized health costs of sugary drink consumption). Therefore, the average value of B across all consumers is proportional to the average demand elasticity times the average externality, plus the *covariance* of the two. This covariance term reflects the fact that if consumers who generate the largest externalities are most responsive to a tax, then the externality benefits of a corrective tax are larger.

The transfer $A + C$ has the same height for all consumers $(p_t - p_0)$, but its width depends on the quantity of sugary drinks consumed by each consumer. Moreover, this summation across consumers is weighted by the difference between their welfare weight and the value of public funds. In theory, the sign of this welfare effect can be either positive or negative, but it will tend to be negative if poorer consumers (those with high welfare weights) tend to purchase more of the externality-producing good, as is the case for sugar-sweetened beverages. The welfare effect of this transfer depends on the level of sugary drink consumption across the income distribution, and not on sugary drink consumption as a share of consumers' income. Thus, this approach accounts for the common concern that sugar-sweetened beverage taxes may be regressive.

Welfare Effects Due to Internalities

In the context of sugar-sweetened beverages, there are two main reasons why consumers might not act in their own best interest. First, consumers may have imperfect information, and thus they may not know how sugar-sweetened beverages can harm their health. Of course, information provision, such as educational campaigns and disclosure requirements, is the direct way to address imperfect information (as studied by, for example, Bollinger, Leslie, and Sorensen 2011; Cantor et al. 2015; Moran and Roberto 2018; Grummon, Taillie, et al. 2019). However, unless these policies fully inform all consumers, there is a role for taxes as a complementary policy tool.

Second, consumers may face problems of self-control and time-inconsistency and thus might underweight the future health costs of consumption of sugar-sweetened beverages relative to how they would like, in the future, to have weighted those costs. There is disagreement as to whether policymakers should respect consumers' "long-run" or "short-run" preferences (Bernheim and Rangel 2009; Bernheim 2016; Bernheim and Taubinsky 2018). A social planner who uses the long-run criterion for welfare analysis might want to help people implement their long-run preferences by reducing consumption of sugar-sweetened beverages.

We can reinterpret Figure 2 to illustrate internalities (assuming away externality costs for the moment), with D_1 representing the consumer's observed demand curve and D_2 representing the latent demand curve that would arise if consumers did not suffer from internalities. Then the vertical distance b represents an ignored internality cost, measured in money units.

Internalities operate similarly to externalities, with one important difference: the area $B + C$ accrues to the *consumer*, rather than to the bearer of the externality (the government in our example above). This does not change the interpretation of the transfer A, from the individual to the government, which will again take on a more negative value if poorer consumers purchase more sugar-sweetened beverages. And the area C can be regarded as a transfer from consumers affected by the tax to themselves, so for social welfare purposes, it can be ignored. However, it does change the interpretation of B, which (unlike in the case of externalities) is multiplied by the individual's welfare weight. As a result, for a given *average* size of the internality, the internality correction benefits from the tax are larger to the extent that poorer consumers have larger areas of B. This will be the case either if internalities are larger for poor consumers (for example, due to poorer access to information or more exposure to settings that demand and deplete self-control) or if their demand response Δq is higher (for example, if the elasticity of demand is constant across consumers, since the poor consume a higher level). In other words, internality benefits from a sugary drink tax are theoretically likely to be progressive, even if the financial costs are regressive.

In a context with both externalities and internalities, one must add the externality and (welfare-weighted) internality benefits, netted against any welfare effects due to the transfer of resources from consumers to the government. Externality benefits depend (positively) on the aggregate elasticity of demand for sugar-sweetened beverages, the average externalized health cost from consumption, and their covariance. Internality benefits similarly depend on the aggregate elasticity and average uninternalized health costs, as well as the extent to which uninternalized health costs and demand responses are higher among poor consumers. Finally, the welfare cost of the resource transfer is larger to the extent that poor households consume more sugary drinks.

Are Sugar-Sweetened Beverage Taxes "Regressive"?

A common concern about sugar-sweetened beverage taxes is that they may hurt poor households, since low earners tend to purchase more sugary drinks. The concepts of externalities, internalities, and transfers from Figure 2 illustrate the basic forces at work.

To understand who is helped and hurt by a sugar-sweetened beverage tax, we need to draw a distinction between *who pays the most in taxes* and *who is benefited or harmed, all things considered*. While it is true that poorer consumers will pay more in taxes on average (due to their higher sugary drink consumption), if there are internality costs from consuming sugary drinks, the beneficial reductions of health conditions such as heart disease and diabetes will also accrue to low-income households, as highlighted by Gruber and Kőszegi (2004). In terms of Figure 2, although poorer consumers incur more costs due to area A on average, those may be offset (partially, or more than fully) by the gained area B. As a result, the fact that poorer consumers purchase more sugar-sweetened beverages does not necessarily imply that they are made worse off by the tax. The extent of this offset depends on the price elasticity of demand: if consumers substantially reduce sugar-sweetened beverage consumption in response to a tax, then the corrective benefits are large relative to the financial burden, making the tax less regressive. On the other hand, if a tax has little effect on consumption, then the corrective benefits are relatively small.

A related question is how the profile of consumption by income affects the optimal *size* of the tax. This depends on why consumption varies with income. Do people have the same underlying preferences, so differences in consumption across incomes are due to the causal effect of more or less income? Or do people at different income levels have systematically different preferences, so that they would consume different amounts even if their incomes were all reset to the same level? A classic principle of optimal taxation (Atkinson and Stiglitz 1976) holds that if differences in consumption of sugary drinks (or any other good) are driven by causal income effects, then they should not be taxed or subsidized for redistributive purposes—such redistribution is more efficiently carried out through the income tax. In contrast, if differences in sugary drink consumption are driven by between-income preference heterogeneity, then that consumption serves as a "tag," which is useful for redistribution, reducing the optimal sugary drink tax. In Allcott, Lockwood, and Taubinsky (2019), we find that preference heterogeneity appears to be the reason why low-income people drink more sugar-sweetened beverages.

Finally, the total regressivity of a sugar-sweetened beverage tax may depend on how the resulting revenues are allocated. Some existing policies have earmarked revenues toward causes that primarily benefit low-income households—sometimes called "progressive revenue recycling." In Philadelphia, for example, a portion of sugar-sweetened beverage tax revenues is pre-allocated to expanding prekindergarten education services within the city. Although earmarking may be useful for building popular support for sugary drink taxes, from a theoretical perspective the practice does not alter the size of the optimal tax. To the extent that it is beneficial to target funds toward prekindergarten programs—or to make the income tax-and-transfer schedule more progressive generally—then that should be done regardless of whether a sugar-sweetened beverage tax is implemented. As a result, revenue recycling and earmarking may be better interpreted as questions of political expediency, rather than optimal taxation. Moreover, pre-allocation may create challenges for policymakers if the tax turns out to be more effective than expected at reducing sugary drink consumption, resulting in a budget shortfall for popular or progressive programs.

Substitution and Leakage

So far, we have assumed that sugar-sweetened beverages can be modeled as one homogeneous good with no substitutes or complements. In reality, this is not the case, which generates additional important considerations.

First, there are many thousands of different sugar-sweetened beverages, each with different sugar content. Theoretically, the optimal structure would be to impose separate taxes on each good, depending on the parameters described above (internalities, externalities, demand elasticities, and between-income preference heterogeneity). In practice, these parameters are difficult to estimate for each specific good, and such heterogeneous taxes would be prohibitively difficult to administer. Most existing sugar-sweetened beverage taxes therefore use a simplified structure of a constant tax rate per ounce of drink. However, since the externalities and internalities from sugary drinks come from the sugar, not the liquid, the externalities and internalities are likely to be proportional to the sugar content of beverages. An alternative simple tax structure of a constant tax rate per gram of sugar in the drink would much more closely approximate the theoretical optimum.

Second, when consumers cut back on sugar-sweetened beverages due to the tax, they may also raise or lower their consumption of other (untaxed) sugary goods.

To the extent that they do, the resulting change in externalities and internalities from those goods should be considered when setting the tax on sugar-sweetened beverages. The sign of this effect is ambiguous. For example, consumers may view sugary snacks as a substitute for sugar-sweetened beverages—an alternative way to get a desired "sugar kick"—in which case some of the internality and externality reductions from a tax on sugar-sweetened beverages may be offset by increased internalities and externalities from substitution to other sugary goods. On the other hand, sugar-sweetened beverages and unhealthy foods may be complements, and if consumers tend to purchase or consume such snacks together, then the analysis above will *understate* the benefits of a sugary drink tax.

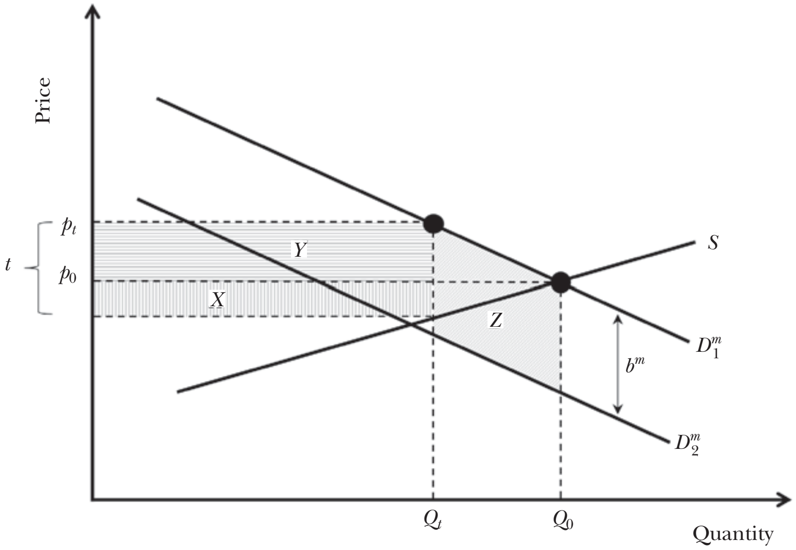
A third reason substitution may matter is that consumers may adjust their behavior to evade or avoid a sin tax—for example, through black market cigarette or drug purchases or, in the case of city-level beverage taxes, through cross-border shopping. This so-called tax "leakage" creates costs for consumers without reducing externalities and internalities from sugar-sweetened beverage consumption. As a result, although some local tax experimentation is useful for estimating the effects of a tax, in the long run there is a benefit from harmonizing tax rates to reducing avoidance by setting them at the state or regional level.

Pass-Through and Producer Surplus

The exposition so far accounts only for the consumer side of the market and therefore leaves out two key issues: the question of tax pass-through (what portion of the tax is borne by consumers in the form of a price increase) and the phenomenon of producer surplus (which accrues to firm owners, in the form of profits, or to employees). To illustrate these forces, Figure 3 depicts a simple supply-and-demand model of the market for sugar-sweetened beverages. D_1^m represents observed market demand for sugar-sweetened beverages, while b^m represents the average marginal externality (weighted by elasticities of demand) plus average marginal internality (weighted by elasticities of demand and welfare weights), so that D_2^m represents market demand less the uninternalized social cost of consumption (normalized by the marginal value of public funds) at each quantity. For illustrative purposes, the pictured tax is a little lower than the optimal level, b^m.

In a simple model like this one, the conventional explanation for incomplete tax pass-through is that some of the tax incidence falls on producers rather than consumers. To account for this possibility, we allow for a market supply curve S that slopes upward, due, for example, to rising marginal costs. The share of the tax that is passed through to consumers is $\frac{p_t - p_0}{t}$, a quantity that rises with the elasticity of sugary drink supply and falls with the (absolute) elasticity of demand. The tax then has three distinct effects on welfare: a transfer from producer surplus to the government, represented by the vertically hatched area X; a transfer from consumers to the government, represented by the horizontally hatched area Y; and a beneficial reduction in externalities and internalities (now combined), represented by the diagonally hatched area Z.

Figure 3
Effect of a Sugar-Sweetened Beverage Tax on Market Consumption



Source: The authors; see Allcott, Lockwood, and Taubinsky (2019) for a formal treatment.
Notes: The figure depicts a simple supply-and-demand model of the market for sugar-sweetened beverages. D_1^m represents observed market demand for sugar-sweetened beverages and b^m the average marginal externality (weighted by elasticities of demand) plus average marginal internality (weighted by elasticities of demand and welfare weights), so that D_2^m represents market demand less the uninternalized social cost of consumption at each quantity. The pictured tax is a little lower than the optimal level, b^m. The area *X* represents a transfer from producers to the government; the area *Y*, a transfer from consumers to the government; and the area *Z*, a beneficial reduction in externalities and internalities (now combined).

Relative to a model with infinitely elastic supply of sugary drinks (corresponding to full pass-through to consumers), the key difference is that some of the costs of the tax are borne by producers rather than consumers. If marginal resources are valued equally in the hands of producers and (welfare-weighted) consumers of sugar-sweetened beverages, the issue of pass-through is irrelevant: in this case, the tax should be adjusted to maximize the welfare gain from the internality and externality benefit *Z*, and the weighted transfer of resources *X* + *Y*. But if resources are valued more in the hands of consumers than producers of sugar-sweetened beverages—for example, if marginal resources accrue to firm shareholders who have a lower average welfare weight than consumers of sugar-sweetened beverages (perhaps because they have higher incomes)—then a lower pass-through will imply a larger net welfare benefit from the tax and a higher tax at the optimum. Conversely, if a higher welfare weight is placed on producers, then partial pass-through calls for a lower optimal sugar-sweetened beverage tax.

Other explanations for partial pass-through, such as discrete pricing policies by grocers or an inability to separately price regular and diet soda fountain sales at fast-food restaurants, might generate different implications. In particular, if a portion of the tax is absorbed by producers with no reduction in quantity supplied, then the optimal tax may need to be larger than b^m to achieve the efficient reduction in sugary drink consumption. However, this possibility depends on understanding the reason for partial pass-through, in addition to quantifying the pass-through rate itself.

Empirical Estimates of Key Parameters

In this section, we review the empirical estimates of the key parameters identified in the previous section, with an eye to the strengths and weaknesses of different estimation strategies.

Demand Elasticities

When estimating demand for any good, not just sugar-sweetened beverages, perhaps the most basic challenge is to isolate quasi-random price variation in order to estimate the demand curve. Conceptually, a demand curve reflects the causal effect of prices on quantity purchased, not just the correlation between prices and quantity purchased. The ideal way to estimate a demand curve would be to run an experiment in which different consumers are offered different prices and then to measure the share of consumers that buy at each price. When market data do not include randomized pricing experiments, several factors will mean that correlation doesn't imply causation. For example, measurement error in prices can also incorrectly make demand appear to be less responsive to price than it actually is. As another example, retailers naturally charge higher prices for higher-quality goods, as well as higher prices for the same good in periods of high demand. This "simultaneity bias" can sometimes even generate positive correlations between price and quantity demanded, whereas the true causal relationship is negative.

There are two types of strategies for isolating quasi-random variation in nonexperimental data. The first is to attempt to control for product quality and demand fluctuations, in hopes that the remaining price variation is quasi-random. For example, Dubois, Griffith, and O'Connell (2017) include brand, time, and other fixed effects, thereby identifying the demand elasticity only off of variation in prices of the same product across retailers and variation in the slope of nonlinear pricing (the relative prices of small versus large containers) across brands. The second strategy is to find a useful instrumental variable for exogenous price movements. In Allcott, Lockwood, and Taubinsky (2019), we create an index of the price households pay for the specific sugar-sweetened beverages they buy at the specific stores where they shop, and we instrument for that price with the time-varying prices that the same retailer charges for the same beverages at other stores in other counties. Finkelstein et al. (2013) instrument for a household's price paid with prices paid by other households in the same city and quarter, excluding households living in the household's Census tract.

For sugar-sweetened beverages, data availability is a particular challenge. There are two common types of datasets. The first is household-level scanner data, such as the US National Consumer Panel (also known as Nielsen Homescan) or Kantar Worldpanel. Participating households are asked to scan the bar codes of all groceries that they bring home, but they do not record consumption away from home, such as purchases at restaurants, vending machines, and ballparks. This unobserved consumption can be substantial: in Allcott, Lockwood, and Taubinsky (2019), we estimate that total consumption exceeds Homescan grocery purchases by 39 percent. If sugar-sweetened beverage taxes are imposed on all consumption, then the relevant demand elasticity is for all consumption, including away from home. Consumption away from home could be more or less price elastic, and there may also be bias due to substitution if households respond to higher grocery prices by consuming more away from home. The second type of dataset is self-reported consumption from beverage frequency questionnaires or dietary recall studies such as the National Health and Nutrition Examination Survey, in which people record food and drink consumed over the past 24 hours or some other recent period. Self-reports may have more measurement error and do not track the same individuals over time, making it difficult to use the two strategies detailed above for isolating quasi-exogenous price variation.

There are several reviews of the literature estimating the price elasticity of demand for sugar-sweetened beverages. Andreyeva, Long, and Brownell (2010) report that across 14 studies, the mean price elasticity is −0.79, with a range from −0.13 to −3.18. Powell et al. (2013) review 12 studies and find a mean price elasticity of −1.21, with a range from −0.71 to −3.87. In Allcott, Lockwood, and Taubinsky (2019), we estimate an elasticity of about −1.4. This relatively elastic demand implies that the internality and externality reduction benefits from a tax are meaningful relative to the burden of the tax payments.

A separate but related parameter is the elasticity of sugar-sweetened beverage consumption with respect to a *tax*. As illustrated in Figure 3, the tax elasticity depends on both the supply and demand elasticities. The tax elasticity is of interest because it determines the public health effect of a tax. Fletcher, Frisvold, and Tefft (2010) study how consumption of sugar-sweetened beverages responds to changes in whether they are included in state sales and excise taxes, but this variation is very limited: among states with a nonzero tax during their sample period, the average tax rate was no more than about 5 percent. Bollinger and Sexton (2017), Cawley, Frisvold, et al. (2018b), Silver et al. (2017), Seiler, Tuchman, and Yao (2019), and others study responses to the Berkeley and Philadelphia taxes. While these tax rates are higher than the taxes studied by Fletcher, Frisvold, and Tefft (2010), having only one or two cities limits the sample size and requires the strong assumption that no factors other than the tax change affected sugar-sweetened beverage demand. Tax elasticity estimates may also capture how interest groups' advertising campaigns and public debates about sin taxes could affect demand over and above the effect of a price increase (Taylor et al. 2016; Rees-Jones and Rozema 2018).

Externalities

Sugar-sweetened beverage consumption generates two main types of externalities: health cost externalities and other fiscal externalities. Estimating the size of these externalities involves a series of challenges in measurement and causal inference.

Health cost externalities result because most Americans have health insurance, typically through their employers, Medicare, or Medicaid, and thus most of the health costs caused by sugar-sweetened beverage consumption are paid for by others. Wang et al. (2012) and Long et al. (2015) both estimate that the health system costs of sugar-sweetened beverages are approximately 1 cent per ounce of sugar-sweetened beverage consumed. The US Department of Health and Human Services estimates that for people with employer-provided insurance, about 15 percent of health costs are borne by the individual, while 85 percent are covered by insurance (Yong, Bertko, and Kronick 2011). Cawley and Meyerhoefer (2012) estimate that 88 percent of the total medical costs of obesity are borne by third parties. Putting these numbers together suggests that the average health cost externality from sugar-sweetened beverage consumption might be 0.8 to 0.9 cents per ounce.

This figure might overstate the true externality, because the results of Bhattacharya and Bundorf (2009; see also Bhattacharya and Sood 2011 in this journal) suggest that obese people who have employer-sponsored health insurance face the full health costs of obesity through lower wages. However, it is not clear whether these labor market effects also exist for less easily observable diseases such as diabetes and cardiovascular disease, and the results do not apply to people with government-sponsored health insurance through Medicare or Medicaid.

In addition to health cost externalities, sugar-sweetened beverage consumption imposes other fiscal externalities, like positive or negative effects on the government's budget. As one tragic example, obesity appears to cause people to die earlier, reducing the amount of Social Security benefits that obese people will claim (Fontaine et al. 2003; Bhattacharya and Sood 2011).

As described in the previous section, the key statistic is the average externality from sugar-sweetened beverage consumption for people who respond to a small change in the tax. While we have estimates of average externalities and overall demand elasticity, one additional important but unknown statistic is the covariance across people between the demand elasticity and the marginal health damages of sugar-sweetened beverage consumption. For example, low-income people are thought to be more price elastic, and their health cost externalities may be higher (if their health costs are not offset by wage reductions because they are on Medicaid) or lower (if they are more likely to be uninsured). Dubois, Griffith, and O'Connell (2017) argue that sugar-sweetened beverage consumption by young people might generate larger health harms, and they show that young people are more price elastic. Sugar-sweetened beverage consumption by people who are prediabetic—that is, just below the threshold for receiving diabetes treatment—may generate larger health cost externalities, since additional consumption may result in high health costs from managing type 2 diabetes. This covariance is one of many questions for future research.

Internalities

As with externalities, there are multiple challenges to measuring internalities.[2] First, there is a mechanical tension in evaluating policies to address internalities, which are predicated on the idea that consumers *do not* act in their own best interest, using revealed preference techniques, which are predicated on the idea that consumers *do* act in their own best interest. Following Bernheim and Rangel (2009), behavioral welfare analyses must somehow establish a "welfare-relevant domain"—that is, a subset of consumer choices that are assumed to be unbiased—versus another subset of "suspect" choices that may be affected by bias. This requires assumptions. Second, measuring internalities often involves the same type of causal inference challenges that arise when estimating price elasticities, health effects, and other parameters. Third, internalities must be measured in units of dollars, as highlighted by the fact that the internality and/or externality b is a vertical distance separating the demand curves in Figures 2 and 3. While much of the behavioral economics literature has focused on simply establishing the presence of some behavioral bias, behavioral welfare analysis requires that internalities be quantified in units of dollars.

As discussed above, imperfect information and lack of self-control are two primary reasons why consumers might not act in their own best interest. Different empirical strategies are often required to quantify different types of internalities. For imperfect information, researchers can estimate the effects of information provision, as in Allcott and Taubinsky (2015) and others. For self-control, researchers can compare choices made for consumption now versus in the future, as in Read and van Leeuwen (1998), Augenblick, Niederle, and Sprenger (2015), and others. For example, Sadoff, Samek, and Sprenger (2015) take advance orders for grocery delivery and allow people to re-optimize their choices at the time that the groceries are delivered, finding that people tend to re-optimize toward less-healthy options and that one-third of people would like to restrict their own future ability to re-optimize. However, standard "preference reversal" experiments cannot directly quantify the effects of limited self-control in dollar units.[3]

Alternatively, a "counterfactual normative consumer" approach can be used to measure multiple biases simultaneously, and to quantify their effects in dollar terms. As an example of this approach, Bronnenberg et al. (2015) show that sophisticated shoppers—in their application, doctors and pharmacists—are more likely to buy generic instead of branded drugs, and they conduct welfare analysis assuming that only sophisticates' choices are welfare relevant. Bartels (1996), Handel and Kolstad

[2] A growing literature in behavioral economics attempts to measure bias in various settings: for overviews, see Allcott and Sunstein (2015), Bernheim and Rangel (2009), Bernheim and Taubinsky (2018), DellaVigna (2009), Handel and Schwartzstein (2018), and Mullainathan, Schwartzstein, and Congdon (2012).

[3] Another approach to quantifying self-control problems is to combine an outside estimate of time-inconsistency from another domain with an estimate of the future private costs of sugar-sweetened beverage consumption. However, it is difficult to assess those future private costs, and the extent of time-inconsistency can vary across domains.

(2015), Johnson and Rehavi (2016), and Levitt and Syverson (2008) similarly compare informed to uninformed agents to identify the effects of imperfect information.

In Allcott, Lockwood, and Taubinsky (2019), we use this counterfactual normative consumer approach to measure the effect of both imperfect information and self-control on sweetened beverage consumption. Specifically, we survey Nielsen Homescan panelists to measure nutrition knowledge and perceived overconsumption of sugar-sweetened beverages, and we find that soda consumption is higher among consumers who are less informed about nutrition and who profess less self-control, even after controlling for demographic variables and survey-based measures of health preferences and tastes for different drinks. The key weakness of this approach is that it requires the assumption that the conditional correlation between bias and consumption equals the causal effect of bias on consumption. Under this assumption, we predict that the average American household would consume 31 to 37 percent less sugar-sweetened beverage if they had perfect self-control and had the nutrition knowledge of dietitians and nutritionists. Translated into dollar terms, the estimated average marginal internality from sugar-sweetened beverage consumption is 0.91 to 2.14 cents per ounce.

Regressivity

The progressivity or regressivity of a sin tax depends on how the internality-reduction benefits and the burden of tax payments vary across the income distribution. In Allcott, Lockwood, and Taubinsky (2019), we find that internality-reduction benefits are highly progressive. Lower-income people have systematically less nutrition knowledge and are more likely to self-report that they consume more sugary drinks than they think they should. While not dispositive, these facts suggest that lower-income people have larger internalities than higher-income people. Our estimated average marginal internality is about one-third larger at household incomes below $10,000 per year compared with at household incomes above $100,000 per year. Furthermore, low-income households reduce sugar-sweetened beverage consumption much more than high-income households when prices rise. Specifically, we find very similar price elasticities, so high-income and low-income households reduce consumption by similar *proportions* in response to a price increase. But because low-income households consume much more, the absolute amounts of their reductions are much larger. Putting these facts together implies that internality-reduction benefits are highly progressive. Under conventional degrees of inequality aversion used in models of optimal income taxation, this progressivity magnifies the internality correction in the optimal tax formula by about 20 percent.

On the other hand, because low-income households consume more sugar-sweetened beverages, they pay more in tax payments. Combining the progressivity of internality-reduction benefits with the regressivity of the tax payments, we find that the net benefits of a sugar-sweetened beverage tax are reasonably flat across the income distribution, and are possibly highest for the lowest-income consumers. More importantly, we find that low-income people benefit substantially from sugar-sweetened beverage taxes, regardless of whether they benefit more than high-income people.

Substitution and Leakage

As described above, the welfare effects of sugar-sweetened beverage taxes depend on whether they affect consumption of other untaxed goods that generate externalities or internalities. Various papers estimate demand systems that capture these substitution patterns between sugar-sweetened beverages and other foods and beverages. Possibly due to the challenges in data quality and variation in identification strategies, there is very little agreement in this literature. For example, Duffey et al. (2010) find that pizza is a strong substitute for sugar-sweetened beverages. Finkelstein et al. (2013) find no substitution to pizza, but statistically significant substitution to canned soup. Zhen et al. (2014) find that canned soup is a complement to carbonated soft drinks but a substitute for sports drinks, energy drinks, and juice drinks. Aguilar et al. (2019) use the implementation of beverage and food taxes in Mexico to estimate substitution to untaxed goods. These conflicting and sometimes counterintuitive results highlight the difficulties in estimating substitution patterns. They may also reflect false positives from multiple hypothesis testing, as there is not an obvious reason for why pizza and canned soup are substitutes for sugary drinks.

In Allcott, Lockwood, and Taubinsky (2019), we find that diet drinks are moderate substitutes for sugar-sweetened beverages. Across a comprehensive range of other drink categories, sugary foods, and even alcohol and cigarettes, we find close to zero net substitution from sugar-sweetened beverages to other possible "sin goods." This would imply that welfare evaluations and optimal tax calculations could safely ignore substitution to other goods, unless one believes that diet drinks have material health harms.

In addition to substitution to other goods, evaluations of local taxes also need to account for substitution to sugar-sweetened beverages purchased outside of the taxed jurisdiction. Bollinger and Sexton (2017) find that approximately half of purchase reductions of sugar-sweetened beverages within Berkeley appear to be substituted to retailers just outside of Berkeley. Roberto et al. (2019) and Seiler, Tuchman, and Yao (2019) also detect substitution to purchases outside of Philadelphia in response to the Philadelphia tax. This leakage reduces the welfare gains from the city-level taxes and reduces the optimal tax rate.

Pass-Through and Producer Surplus

To ease administration and to increase tax salience, city-level sugar-sweetened beverage taxes in the United States are generally collected from beverage distributors that sell to retailers. A number of recent papers have estimated the extent to which these taxes are passed through into higher retail prices. Two papers studying the Philadelphia tax conclude that the tax was approximately fully passed through (Cawley, Frisvold, et al. 2018a; Seiler, Tuchman, and Yao 2019). Six papers studying Berkeley and Boulder find less than full pass-through, implying that at least some of the incidence of these taxes is on suppliers (Falbe et al. 2015; Bollinger and Sexton 2017; Cawley and Frisvold 2017; Rojas and Wang 2017; Silver et al. 2017; Cawley, Crain, et al. 2018).

Bollinger and Sexton (2017) also document how retailers' overall pricing strategies interact with a local tax on a small subset of products. First, as documented by DellaVigna and Gentzkow (2017) and Hitsch, Hortacsu, and Lin (2017), large retail chains often set uniform prices across many stores in many cities. This limits the extent to which a local cost increase from a local tax is passed through to retail prices in that area. Second, retailers often use "category pricing": for example, all two-liter bottles of regular *and* diet soda might have the same price. If retailers maintain equal prices for regular and diet soda and if consumption of diet soda involves lower internalities or externalities because it does not contain sugar, this reduces the welfare gains from a tax on sugar-sweetened beverages. This intersection between industrial organization and optimal taxation is an interesting area for further research.

Putting It Together

In Allcott, Lockwood, and Taubinsky (2019), we estimate that the socially optimal sugar-sweetened beverage tax is between 1 and 2.1 cents per ounce. One can understand this as coming from the correction needed to offset the negative externality (about 0.8 cents per ounce) and internality (about 1 cent per ounce, inflated by 20 percent due to the progressivity of internality correction), with a further reduction due to the regressive incidence of the financial costs of a tax (reducing the tax by about 0.5 cents per ounce). Together, these rough estimates suggest an optimal tax of about 1.5 cents per ounce. While there is considerable uncertainty in these optimal tax estimates, the optimal tax is not zero and may be higher than the levels in most US cities to date. However, for policymakers who are philosophically opposed to considering internalities in an optimal tax calculation, the optimal tax considering only externalities is around 0.4 cents per ounce.

Guiding Principles for Policymakers

Although uncertainty remains about some empirical parameters, economic theory and existing data suggest seven guiding principles for designing sugar-sweetened beverage taxes. The first four principles are all motivated by one deeper principle: that sin taxes should be designed to offset uninternalized harms.

1. Focus on Counteracting Externalities and Internalities, Not on Minimizing Sugary Drink Consumption

Many public health advocates explicitly or implicitly take the perspective that the goal of policymakers should be to maximize health or minimize unhealthy behaviors. It's easy to see why this can't be the right social objective. The way to maximize health is to ban any sugary or fatty food or drink, including sugary drinks, red meat, and dessert. Such a ban would preclude any enjoyment that people get from eating steak or dessert, and it's not clear where to draw the line on what foods or drinks to ban.

The economic framework presented in this article instead focuses on maximizing social welfare and provides a principled approach that trades off health-related externalities and internalities with consumer surplus, producer surplus, and government revenues. The framework highlights that unhealthy behaviors do not necessarily merit policy intervention, as they could simply reflect the fact that people enjoy eating steak and dessert. Sin taxes are justified only to the extent that they offset uninternalized externalities or internalities.

2. Target Policies to Reduce Consumption among People Generating the Largest Externalities and Internalities

Consumption by different people may involve larger or smaller externalities and internalities, perhaps due to differences in self-control, nutrition knowledge, and health insurance coverage. Ideally, policies would be targeted to reduce consumption more among people with larger externalities and internalities. For example, if internalities and externalities are largest among children—perhaps due to limited self-control, or because their consumption generates lifelong habits—then very high taxes or bans on sugar-sweetened beverages in schools may be justified.

3. Tax Grams of Sugar, Not Ounces of Liquid

Most sugar-sweetened beverage taxes are structured as a per-ounce tax on any drink with added sugar. That means that drinks with high and low amounts of added sugar are taxed at the same rate. From the perspective of the theoretical rationale for sugary drink taxes, this structure makes little sense. It's the sugar in the drinks, not the amount of liquid, that harms our health. Therefore, drinks containing more sugar generate greater externalities and probably greater internalities.

Scaling the tax with the amount of sugar instead of the amount of liquid that comes with the sugar encourages consumers to switch to lower-sugar drinks and also encourages producers to reduce sugar content. Using economic and epidemiological models, we estimate that taxing sugar-sweetened beverages based on sugar content instead of volume would boost a tax's health benefits by 43 percent, helping people around the world to lose nearly 200 million pounds (Grummon, Lockwood, et al. 2019). Other research arrives at qualitatively similar conclusions about the gains from taxing sugar content instead of volume (Francis, Marron, and Reuben 2016; Zhen, Brissette, and Ruff 2014).[4]

4. Tax Diet Drinks and Fruit Juice If and Only If They Also Cause Uninternalized Health Harms

Even if restricted to 1 cent per ounce volumetric taxes, policymakers must decide what drinks should be included in sugary drink taxes. The Philadelphia and erstwhile Cook County taxes also include diet drinks, on the grounds that this raises more revenues and also makes the tax less regressive because higher-income people buy more diet drinks. However, the Philadelphia diet drink tax is an inefficient way

[4] The United Kingdom and several other countries approximate sugar taxes through tiered systems that impose a higher volumetric tax for drinks with higher sugar content, but this still falls short of the ideal of setting taxes proportional to uninternalized harms.

to raise revenue, and as we discuss below, the regressivity argument is misguided. Including diet drinks would be justified only if the externalities and internalities from diet drinks are as large as those from nondiet drinks, but the evidence presented above suggests that diet drinks are less harmful.

All existing sugary drink taxes exclude 100 percent fruit juice, despite arguments by Wojcicki and Heyman (2012), Gill and Sattar (2014), and some other public health experts that the naturally occurring sugar in fruit juice may be as harmful as the added sugar in soft drinks. Exempting fruit juice from a beverage tax is justified only if the positive externalities and internalities from the additional vitamins and nutrients offset the negative externalities and internalities from the sugar.

5. When Judging Regressivity, Consider Internality Benefits, Not Just Who Pays the Taxes

Some people argue that sugar-sweetened beverage taxes are regressive, because low-income people buy more of these beverages and will thus pay more in taxes. As we discussed above, however, what matters is not just how much low-income people would pay in this kind of tax but how much this tax benefits or harms them overall. In Allcott, Lockwood, and Taubinsky (2019), we estimate that low-income people enjoy a disproportionate share of the internality-reduction benefits, because they both have larger internalities in this domain and reduce consumption more in response to a tax. Overall, our results suggest that low-income people benefit substantially from sugar-sweetened beverage taxes, and they may even benefit more than high-income people.

6. If Possible, Implement Taxes Statewide

All of the current sugar-sweetened beverage taxes in the United States have been implemented by individual cities. Evidence suggests that the benefits of city-level taxes are diminished because consumers avoid these taxes by purchasing outside of the city. To reduce this leakage, sugar-sweetened beverage taxes would ideally be implemented over larger geographic areas, such as at the state level. Such geographic integration can also help to reduce the importance of compliance and administrative costs. However, the existence of externalities and internalities suggests that city-level taxes may be better than no taxes at all.

7. The Benefits of Sugar-Sweetened Beverage Taxes Probably Exceed Their Costs

Our read of the evidence is that sugar-sweetened beverage consumption likely imposes externalities on the health system and internalities due to imperfect nutrition knowledge and self-control problems. In Allcott, Lockwood, and Taubinsky (2019), we estimate that the social welfare benefits from implementing the optimal tax nationwide (relative to having zero tax) are between $2.4 billion and $6.8 billion per year. These gains would be substantially larger if the tax rate were to scale with sugar content.

Of course, such calculations require strong assumptions and depend on uncertain empirical estimates, in particular with respect to internalities and externalities.

We therefore emphasize that much more empirical work is needed. Furthermore, sugar-sweetened beverage taxes are not a panacea—they will not, by themselves, solve the obesity epidemic in America or elsewhere. But sin taxes have proven to be a feasible and effective policy instrument in other domains, and the evidence suggests that the benefits of sugar-sweetened beverage taxes likely exceed the costs.

■ *We are grateful to David Frisvold, Anna Grummon, Gordon Hanson, Robert Inman, Daniel Reck, Alex Rees-Jones, Christina Roberto, Steven Sexton, and Timothy Taylor for helpful comments, and to Andrew Joung for excellent research assistance. We thank the Sloan Foundation and the Wharton Dean's Research Fund for grant funding.*

References

Aguilar, Arturo, Emilio Gutierrez, and Enrique Seira. 2019. "The Effectiveness of Sin Food Taxes: Evidence from Mexico." Instituto Tecnológico Autónomo de México (ITAM) Working Paper.

Allcott, Hunt, Benjamin B. Lockwood, and Dmitry Taubinsky. 2019. "Regressive Sin Taxes, with an Application to the Optimal Soda Tax." *Quarterly Journal of Economics* 134(3). https://doi.org/10.1093/qje/qjz017.

Allcott, Hunt, and Cass R. Sunstein. 2015. "Regulating Internalities." *Journal of Policy Analysis and Management* 34(3): 698–705.

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105(8): 2501–38.

Andreyeva, Tatiana, Michael W. Long, and Kelly D. Brownell. 2010. "The Impact of Food Prices on Consumption: A Systematic Review of Research on the Price Elasticity of Demand for Food." *American Journal of Public Health* 100(2): 216–22.

Atkinson, Anthony B., and Joseph E. Stiglitz. 1976. "The Design of Tax Structure: Direct versus Indirect Taxation." *Journal of Public Economics* 6(1–2): 55–75.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. "Working over Time: Dynamic Inconsistency in Real Effort Tasks." *Quarterly Journal of Economics* 130(3): 1067–115.

Bartels, Larry M. 1996. "Uninformed Votes: Information Effects in Presidential Elections." *American Journal of Political Science* 40(1): 194–230.

Bernheim, B. Douglas. 2016. "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics." *Journal of Benefit-Cost Analysis* 7(1): 12–68.

Bernheim, B. Douglas, and Antonio Rangel. 2009. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *Quarterly Journal of Economics* 124(1): 51–104.

Bernheim, B. Douglas, and Dmitry Taubinsky. 2018. "Behavioral Public Economics." Chap. 5 in *The Handbook of Behavioral Economics*, vol. 1, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson. Amsterdam: Elsevier.

Bhattacharya, Jay, and M. Kate Bundorf. 2009. "The Incidence of the Healthcare Costs of Obesity." *Journal of Health Economics* 28(3): 649–58.

Bhattacharya, Jay, and Neeraj Sood. 2011. "Who Pays for Obesity?" *Journal of Economic Perspectives* 25(1): 139–58.

Bollinger, Bryan, Phillip Leslie, and Alan Sorensen. 2011. "Calorie Posting in Chain Restaurants." *American Economic Journal: Economic Policy* 3(1): 91–128.

Bollinger, Bryan, and Steven Sexton. 2017. "Local Excise Taxes, Sticky Prices, and Spillovers: Evidence from Berkeley's Soda Tax." https://ssrn.com/abstract=3087966.

Bronnenberg, Bart J., Jean-Pierre Dubé, Matthew Gentzkow, and Jesse M. Shapiro. 2015. "Do Pharmacists Buy Bayer? Informed Shoppers and the Brand Premium." *Quarterly Journal of Economics* 130(4): 1669–726.

Cantor, Jonathan, Alejandro Torres, Courtney Abrams, and Brian Elbel. 2015. "Five Years Later: Awareness of New York City's Calorie Labels Declined, with No Changes in Calories Purchased." *Health Affairs* 34(11): 1893–900.

Cawley, John, Chelsea Crain, David Frisvold, and David Jones. 2018. "The Pass-Through of the Largest Tax on Sugar-Sweetened Beverages: The Case of Boulder, Colorado." NBER Working Paper 25050.

Cawley, John, and David E. Frisvold. 2017. "The Pass-Through of Taxes on Sugar-Sweetened Beverages to Retail Prices: The Case of Berkeley, California." *Journal of Policy Analysis and Management* 36(2): 303–26.

Cawley, John, David Frisvold, Anna Hill, and David Jones. 2018a. "The Impact of the Philadelphia Beverage Tax on Prices and Product Availability." NBER Working Paper 24990.

Cawley, John, David Frisvold, Anna Hill, and David Jones. 2018b. "The Impact of the Philadelphia Beverage Tax on Purchases and Consumption by Adults and Children." NBER Working Paper 25052.

Cawley, John, and Chad Meyerhoefer. 2012. "The Medical Care Costs of Obesity: An Instrumental Variables Approach." *Journal of Health Economics* 31(1): 219–30.

de Ruyter, Janne C., Margreet R. Olthof, Jacob C. Seidell, and Martijn B. Katan. 2012. "A Trial of Sugar-Free or Sugar-Sweetened Beverages and Body Weight in Children." *New England Journal of Medicine* 367(15): 1397–406.

DellaValle, Diane M., Liane S. Roe, and Barbara J. Rolls. 2005. "Does the Consumption of Caloric and Non-caloric Beverages with a Meal Affect Energy Intake?" *Appetite* 44(2): 187–93.

DellaVigna, Stefano. 2009. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature* 47(2): 315–72.

DellaVigna, Stefano, and Matthew Gentzkow. 2017. "Uniform Pricing in US Retail Chains." NBER Working Paper 23996.

Dubois, Pierre, Rachel Griffith, and Martin O'Connell. 2017. "How Well-Targeted Are Soda Taxes?" Centre for Economic Policy Research (CEPR) Discussion Paper DP12484.

Duffey, Kiyah J., Penny Gordon-Larsen, James M. Shikany, David Guilkey, David R. Jacobs Jr., and Barry M. Popkin. 2010. "Food Price and Diet and Health Outcomes: 20 Years of the CARDIA Study." *Archives of Internal Medicine* 170(5): 420–26.

Ebbeling, Cara B., Henry A. Feldman, Virginia R. Chomitz, Tracy A. Antonelli, Steven L. Gortmaker, Stavroula K. Osganian, and David S. Ludwig. 2012. "A Randomized Trial of Sugar-Sweetened Beverages and Adolescent Body Weight." *New England Journal of Medicine* 367(15): 1407–16.

Falbe, Jennifer, Nadia Rojas, Anna H. Grummon, and Kristine A. Madsen. 2015. "Higher Retail Prices of Sugar-Sweetened Beverages 3 Months after Implementation of an Excise Tax in Berkeley, California." *American Journal of Public Health* 105(11): 2194–201.

Finkelstein, Eric A., Chen Zhen, Marcel Bilger, James Nonnemaker, Assad M. Farooqui, and Jessica E. Todd. 2013. "Implications of a Sugar-Sweetened Beverage (SSB) Tax When Substitutions to Non-beverage Items Are Considered." *Journal of Health Economics* 32(1): 219–39.

Fletcher, Jason M., David E. Frisvold, and Nathan Tefft. 2010. "The Effects of Soft Drink Taxes on Child and Adolescent Consumption and Weight Outcomes." *Journal of Public Economics* 94(11–12): 967–74.

Flood, Julie E., Liane S. Roe, and Barbara J. Rolls. 2006. "The Effect of Increased Beverage Portion Size on Energy Intake at a Meal." *Journal of the Academy of Nutrition and Dietetics* 106(12): 1984–90.

Fontaine, Kevin R., David T. Redden, Chenxi Wang, Andrew O. Westfall, and David B. Allison. 2003. "Years of Life Lost Due to Obesity." *JAMA* 289(2): 187–93.

Francis, Norton, Donald Marron, and Kim S. Reuben. 2016. *The Pros and Cons of Taxing Sweetened Beverages Based on Sugar Content.* Washington, DC: Urban Institute. https://www.urban.org/research/publication/pros-and-cons-taxing-sweetened-beverages-based-sugar-content.

Fung, Teresa T., Vasanti Malik, Kathryn M. Rexrode, JoAnn E. Manson, Walter C. Willett, and Frank B. Hu. 2009. "Sweetened Beverage Consumption and Risk of Coronary Heart Disease in Women." *American Journal of Clinical Nutrition* 89(4): 1037–42.

Gill, Jason M. R., and Naveed Sattar. 2014. "Fruit Juice: Just Another Sugary Drink?" *Lancet Diabetes and Endocrinology* 2(6): 444–46.

Global Food Research Program. 2019. *Sugary Drink Taxes around the World.* Chapel Hill, NC: Global Food Research Program. http://globalfoodresearchprogram.web.unc.edu/multi-country-initiative/resources/.

Gruber, Jonathan, and Botond Kőszegi. 2004. "Tax Incidence When Individuals Are Time-Inconsistent: The Case of Cigarette Excise Taxes." *Journal of Public Economics* 88(9–10): 1959–87.

Grummon, Anna, Benjamin B. Lockwood, Dmitry Taubinsky, and Hunt Allcott. 2019. "Designing Better Sugary Drink Taxes: How to Lose 200 Million Pounds with One Simple Trick." Unpublished.

Grummon, Anna, Lindsey Smith Taillie,

Shelley D. Golden, Marissa G. Hall, Leah M. Ranney, and Noel Brewer. 2019. "Impact of Sugar-Sweetened Beverage Health Warnings on Beverage Purchases: A Randomized Controlled Trial." Unpublished.

Handel, Benjamin R., and Jonathan T. Kolstad. 2015. "Health Insurance for 'Humans': Information Frictions, Plan Choice, and Consumer Welfare." *American Economic Review* 105(8): 2449–500.

Handel, Benjamin R., and Joshua Schwartzstein. 2018. "Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care?" *Journal of Economic Perspectives* 32(1): 155–78.

Hitsch, Günter J., Ali Hortacsu, and Xiliang Lin. 2017. "Prices and Promotions in U.S. Retail Markets: Evidence from Big Data." Chicago Booth Research Paper 17-18. https://ssrn.com/abstract=2971168.

Imamura, Fumiaki, Laura O'Connor, Zheng Ye, Jaakko Mursu, Yasuaki Hayashino, Shilpa N. Bhupathiraju, and Nita G. Forouhi. 2015. "Consumption of Sugar Sweetened Beverages, Artificially Sweetened Beverages, and Fruit Juice and Incidence of Type 2 Diabetes: Systematic Review, Meta-analysis, and Estimation of Population Attributable Fraction." *BMJ* 351(8018): h3576.

Johnson, Erin M., and M. Marit Rehavi. 2016. "Physicians Treating Physicians: Information and Incentives in Childbirth." *American Economic Journal: Economic Policy* 8(1): 115–41.

Levitt, Steven D., and Chad Syverson. 2008. "Market Distortions When Agents Are Better Informed: The Value of Information in Real Estate Transactions." *Review of Economics and Statistics* 90(4): 599–611.

Long, Michael W., Steven L. Gortmaker, Zachary J. Ward, Stephen C. Resch, Marj L. Moodie, Gary Sacks, Boyd A. Swinburn, Rob C. Carter, and Y. Claire Wang. 2015. "Cost Effectiveness of a Sugar-Sweetened Beverage Excise Tax in the U.S." *American Journal of Preventive Medicine* 49(1): 112–23.

Loughead, Katherine. 2018. "Sales Taxes on Soda, Candy, and Other Groceries, 2018." Tax Foundation FISCAL FACT 598. https://files.taxfoundation.org/20180706104150/Tax-Foundation-FF598-Groceries-Soda-Candy.pdf.

Ludwig, David S. 2002. "The Glycemic Index: Physiological Mechanisms Relating to Obesity, Diabetes, and Cardiovascular Disease." *JAMA* 287(18): 2414–23.

Malik, Vasanti S., An Pan, Walter C. Willett, and Frank B. Hu. 2013. "Sugar-Sweetened Beverages and Weight Gain in Children and Adults: A Systematic Review and Meta-analysis." *American Journal of Clinical Nutrition* 98(4): 1084–102.

Malik, Vasanti S., Barry M. Popkin, George A. Bray, Jean-Pierre Després, Walter C. Willett, and Frank B. Hu. 2010. "Sugar-Sweetened Beverages and Risk of Metabolic Syndrome and Type 2 Diabetes: A Meta-analysis." *Diabetes Care* 33(11): 2477–83.

Mattes, Richard D., James M. Shikany, Kathryn A. Kaiser, and David B. Allison. 2011. "Nutritively Sweetened Beverage Consumption and Body Weight: A Systematic Review and Meta-analysis of Randomized Experiments." *Obesity Reviews* 12(5): 346–65.

Moran, Alyssa J., and Christina A. Roberto. 2018. "Health Warning Labels Correct Parents' Misperceptions about Sugary Drink Options." *American Journal of Preventive Medicine* 55(2): e19–27.

Mourao, Denise M., Josefina Bressan, Wayne W. Campbell, and Richard D. Mattes. 2007. "Effects of Food Form on Appetite and Energy Intake in Lean and Obese Young Adults." *International Journal of Obesity* 31(11): 1688–95.

Mozaffarian, Dariush, Tao Hao, Eric B. Rimm, Walter C. Willett, and Frank B. Hu. 2011. "Changes in Diet and Lifestyle and Long-Term Weight Gain in Women and Men." *New England Journal of Medicine* 364(25): 2392–404.

Mullainathan, Sendhil, Joshua Schwartzstein, and William J. Congdon. 2012. "A Reduced-Form Approach to Behavioral Public Finance." *Annual Review of Economics* 4(1): 511–40.

Pan, An, and Frank B. Hu. 2011. "Effects of Carbohydrates on Satiety: Differences between Liquid and Solid Food." *Current Opinion in Clinical Nutrition and Metabolic Care* 14(4): 385–90.

Pigou, Arthur Cecil. 1920. *The Economics of Welfare.* London: Macmillan.

Popkin, Barry M., and Corinna Hawkes. 2016. "Sweetening of the Global Diet, Particularly Beverages: Patterns, Trends and Policy Responses." *Lancet Diabetes and Endocrinology* 4(2): 174–86.

Powell, Lisa M., Jamie F. Chriqui, Tamkeen Khan, Roy Wada, and Frank J. Chaloupka. 2013. "Assessing the Potential Effectiveness of Food and Beverage Taxes and Subsidies for Improving Public Health: A Systematic Review of Prices, Demand and Body Weight Outcomes." *Obesity Reviews* 14(2): 110–28.

Raben, Anne, Bente K. Møller, Anne Flint, Tatjana H. Vasilaras, A. Christina Møller, Jens Juul Holst, and Arne Astrup. 2011. "Increased Postprandial Glycaemia, Insulinemia, and Lipidemia after 10 Weeks' Sucrose-Rich Diet Compared to an Artificially Sweetened Diet: A Randomised Controlled Trial." *Food and Nutrition Research* 55(1). https://foodandnutritionresearch.net/index.php/fnr/article/view/588.

Read, Daniel, and Barbara van Leeuwen. 1998. "Predicting Hunger: The Effects of Appetite and Delay on Choice." *Organizational Behavior and Human Decision Processes* 76(2): 189–205.

Rees-Jones, Alex, and Kyle Rozema. 2018. "Price Isn't Everything: Behavioral Response around Changes in Sin Taxes." https://ssrn.com/abstract=3205688.

Roberto, Christina A., Hannah G. Lawman, Michael T. LeVasseur, Nandita Mitra, Ana Peterhans, Bradley Herring, and Sara N. Bleich. 2019. "Association of a Beverage Tax on Sugar-Sweetened and Artificially Sweetened Beverages with Changes in Beverage Prices and Sales at Chain Retailers in a Large Urban Setting." *JAMA* 321(18): 1799–810.

Rojas, Christian, and Emily Wang. 2017. "Do Taxes for Soda and Sugary Drinks Work? Scanner Data Evidence from Berkeley and Washington." https://ssrn.com/abstract=3041989.

Sadoff, Sally, Anya Samek, and Charles Sprenger. 2015. "Dynamic Inconsistency in Food Choice: Experimental Evidence from a Food Desert." Becker Friedman Institute for Research in Economics Working Paper 2572821, Center for Economic and Social Research (CESR)-Schaeffer Working Paper 2015-027. https://ssrn.com/abstract=2572821 .

Saez, Emmanuel, and Stefanie Stantcheva. 2016. "Generalized Social Welfare Weights for Optimal Tax Theory." *American Economic Review* 106(1): 24–45.

Santos, Filipe L., Sara S. Esteves, Altamiro da Costa Pereira, William S. Yancy Jr., and José P. L. Nunes. 2012. "Systematic Review and Meta-analysis of Clinical Trials of the Effects of Low Carbohydrate Diets on Cardiovascular Risk Factors." *Obesity Reviews* 13(11): 1048–66.

Schulze, Matthias B., JoAnn E. Manson, David S. Ludwig, Graham A. Colditz, Meir J. Stampfer, Walter C. Willett, and Frank B. Hu. 2004. "Sugar-Sweetened Beverages, Weight Gain, and Incidence of Type 2 Diabetes in Young and Middle-Aged Women." *JAMA* 292(8): 927–34.

Seiler, Stephan, Anna Tuchman, and Song Yao. 2019. "The Impact of Soda Taxes: Pass-Through, Tax Avoidance, and Nutritional Effects." Stanford University Graduate School of Business Research Paper 19-12. https://ssrn.com/abstract=3302335.

Silver, Lynn D., Shu Wen Ng, Suzanne Ryan-Ibarra, Lindsey Smith Taillie, Marta Induni, Donna R. Miles, Jennifer M. Poti, and Barry M. Popkin. 2017. "Changes in Prices, Sales, Consumer Spending, and Beverage Consumption One Year after a Tax on Sugar-Sweetened Beverages in Berkeley, California, US: A Before-and-After Study." *PLOS Medicine* 14(4): e1002283.

Taylor, Rebecca, Scott Kaplan, Sofia B Villas-Boas, and Kevin Jung. 2016. "Soda Wars: Effect of a Soda Tax Election on Soda Purchases." University of California, Berkeley, Department of Agricultural and Resource Economics (CUDARE) Working Paper Series. https://escholarship.org/uc/item/0q18s7b7.

Te Morenga, Lisa A., Alex J. Howatson, Rhiannon M. Jones, and Jim Mann. 2014. "Dietary Sugars and Cardiometabolic Risk: Systematic Review and Meta-Analyses of Randomized Controlled Trials of the Effects on Blood Pressure and Lipids." *American Journal of Clinical Nutrition* 100(1): 65–79.

Vartanian, Lenny R., Marlene B. Schwartz, and Kelly D. Brownell. 2007. "Effects of Soft Drink Consumption on Nutrition and Health: A Systematic Review and Meta-analysis." *American Journal of Public Health* 97(4): 667–75.

Wang, Y. Claire, Pamela Coxson, Yu-Ming Shen, Lee Goldman, and Kirsten Bibbins-Domingo. 2012. "A Penny-per-Ounce Tax on Sugar-Sweetened Beverages Would Cut Health and Cost Burdens of Diabetes." *Health Affairs* 31(1): 199–207.

Wojcicki, Janet M., and Melvin B. Heyman. 2012. "Reducing Childhood Obesity by Eliminating 100% Fruit Juice." *American Journal of Public Health* 102(9):1630–33.

Xi, Bo, Yubei Huang, Kathleen Heather Reilly, Shuangshuang Li, Ruolong Zheng, Maria T. Barrio-Lopez, Miguel A. Martinez-Gonzalez, and Donghao Zhou. 2015. "Sugar-Sweetened Beverages and Risk of Hypertension and CVD: A Dose–Response Meta-analysis." *British Journal of Nutrition* 113(5): 709–17.

Yong, Pierre L., John Bertko, and Richard Kronick. 2011. *Actuarial Value and Employer-Sponsored Insurance.* Assistant Secretary for Planning and Evaluation (ASPE) Research Brief. Washington, DC: US Department of Health and Human Services.

Zhen, Chen, Ian F. Brissette, and Ryan Richard Ruff. 2014. "By Ounce or By Calorie: The Differential Effects of Alternative Sugar-Sweetened Beverage Tax Strategies." 96(4): 1070–83.

Zhen, Chen, Eric A. Finkelstein, James Nonnemaker, Shawn Karns, and Jessica E. Todd. 2014. "Predicting the Effects of Sugar-Sweetened Beverage Taxes on Food and Beverage Demand in a Large Demand System." *American Journal of Agricultural Economics* 96(1): 1–25.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 228–242

Retrospectives
Lord Keynes and Mr. Say: A Proximity of Ideas

Alain Béraud and Guy Numa

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please contact Joseph Persky, Professor of Economics, University of Illinois, Chicago, at jpersky@uic.edu.

Introduction

In a keynote address delivered in April 2014, former US Treasury Secretary and top White House economic adviser Lawrence Summers (2014) declared, "Jean-Baptiste Say, the patron saint of Chicago economists, enunciated the doctrine in the nineteenth century that supply creates its own demand. . . . It was Keynes's great contribution to explain that was wrong, that in a world where the demand could be for money and for financial assets, there could be a systematic shortfall in demand." Summers expressed an opinion shared by many modern economists and textbooks, whether they are proponents or critics of Say's Law: Keynes was supposedly the anti-Say *par excellence.*

■ *Alain Béraud is Professor of Economics and Associate at THEMA (Théorie Economique, Modélisation et Applications), both at the University of Cergy-Pontoise, Cergy-Pontoise, France. Guy Numa is Assistant Professor of Economics, Colorado State University, Fort Collins, Colorado. Numa is the corresponding author at guy.numa@colostate.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.228 doi=10.1257/jep.33.3.228

In the history-of-thought specialist literature, it has been accepted for some time that Keynes distorted the messages of Say (Baumol 1977, 1999; Jonsson 1997; Clower 2004). However, previous commentators have typically focused on what has come to be called "Say's Law," which Say called "the law of outlets." In this essay, we first assess the arguments used by Keynes to attack Say's system, and we find his criticisms to be ill founded. In doing so, we discuss how to interpret Say's law of outlets.

We then contrast and compare Keynes and Say on some other related topics: demand deficiency, the role of money in the economy, and government intervention. English-speaking readers have often missed out on crucial aspects of Say's thinking on these subjects, particularly regarding monetary matters, because much of Say's work has never been translated into English. For example, Say wrote two multivolume magnum opus works, *Traité d'économie politique* (Say [1803, 1814, 1817, 1819, 1826, 1841] 2006; henceforth *Traité*) and *Cours complet d'économie politique pratique* (Say [1828–29] 2010; henceforth *Cours*), and other lesser-known texts. *Traité* was published in six editions, with significant changes from one edition to the next. Only the fourth edition, published in 1819, was translated into English. *Cours*, along with other lesser-known texts such as *Leçons d'économie politique*, has mostly not been translated into English.[1] Indeed, it seems likely that when Keynes criticized Say, he was relying on the interpretation of Say's work by John Stuart Mill. Apparently unbeknownst to Keynes, Say adopted proto-Keynesian views on several key issues. We do not argue that Keynes's analysis is strictly similar to Say's, but we do identify some similarities and potential sources of agreement between them. Our conclusion is that there is a proximity of ideas between Keynes and Say, two of the most influential figures in the economics discipline, who are too often portrayed as polar opposites.

Thumbnail Biographies

Jean-Baptiste Say (1767–1832) and John Maynard Keynes (1883–1946) came from different traditions and different times. Keynes was born into a Victorian family, trained in mathematics at Cambridge, and spent much of his career in public service in the India Office and the British Treasury (Moggridge 1992; Skidelsky 2005). Say was born in a Huguenot family. His life was very much that of an intellectual in troubled times (Palmer 1997), enduring the political and economic chaos of the French Revolution and of the Napoleonic wars. Say was, successively, a bank clerk, soldier, publicist, managing editor, government official, and factory owner and ultimately became one of the first professors of political economy in France (Blanc and Tiran 2003; Schoorl 2013). Say's profound belief in progress was that of a positivist thinker, revolutionary, and republican writer.

Several similarities between Keynes's and Say's respective trajectories are noteworthy. First, they both developed a keen interest in literary activities and belonged

[1] Jacoud (2013) offers a selection of Say's monetary writings in the English language.

to a group of artists and intellectuals that influenced their thinking. Keynes was influenced by his father, John Neville Keynes; by Bertrand Russell's thought; and by the utilitarian philosophy and morale of George Edward Moore. He joined the Bloomsbury group, an association of bohemian writers and artists. Say, on the other hand, was affiliated with the *Idéologistes*—called *Idéologues* by Napoleon—a group of liberal intellectuals who coalesced around the moral philosophy of the Marquis de Condorcet and Antoine Destutt de Tracy, the work in physiology by Pierre Jean George Cabanis, and the sensualism of Étienne Bonnot de Condillac (Lutfalla 1991; Forget 1999; Schoorl 2013).

Moreover, Keynes and Say both demonstrated a preference for practical economic policy. The majority of Keynes's publications showed concern for practical policy problems and for the empirical aspects of these problems (Patinkin 1976, 14). Practical political economy was also a key feature of Say's thinking (Steiner 1990; Potier 2010, pp. xxx–xxxi; Numa forthcoming [a]). Political economy, Say believed, should promote useful knowledge and tools for managing private and public affairs (Say [1814] 2006, 12n2).[2] The relationship between private and public affairs and the use of economic knowledge as a guide for action seem to be common features of Keynes's and Say's respective systems.

Finally, it is interesting to note that both Keynes and Say edited journals, albeit in different areas. From 1911 to 1945, Keynes edited the *Economic Journal*, published by the Royal Economic Society. Between 1794 and 1800, Say edited *La Décade philosophique, littéraire et politique*, the journal published by the *Idéologistes*, while also directing its printing house.

Keynes's Criticism of Say

John Maynard Keynes rarely refers to Jean-Baptiste Say in *The General Theory of Employment, Interest and Money* (Keynes [1936] 1973; henceforth *General Theory*), giving him only three passing mentions (in chapters 2, 3, and 23). However, in the preface of the French edition of the *General Theory*, Keynes ([1939] 1973, p. xxxv) apparently believed he could not address a French audience without bringing up Say:

> I believe that economics everywhere up to recent times has been dominated, much more than has been understood, by the doctrines associated with the name of J.-B. Say. It is true that his "law of markets" has been long abandoned by most economists; but they have not extricated themselves from his basic assumptions and particularly from his fallacy that demand is created by supply. Say was implicitly assuming that the economic system was always operating up to its full capacity, so that a new activity was always in substitution for,

[2] For instance, in 1819, Say and trader Vital Roux cofounded the world's first business school, the *École Spéciale de Commerce* in Paris—now known as ESCP Europe (Kaplan 2014, 530).

> and never in addition to, some other activity. Nearly all subsequent economic theory has depended on, in the sense that it has required, this same assumption. Yet a theory so based is clearly incompetent to tackle the problems of unemployment and of the trade cycle. Perhaps I can best express to French readers what I claim for this book by saying that in the theory of production it is a final break-away from the doctrines of J.-B. Say and that in the theory of interest it is a return to the doctrines of Montesquieu.

In this passage, Keynes criticizes Say's "law of markets," which Say actually called the "*théorie des débouchés*" and which might be better translated as the "law of outlets"[3]—on three points: (1) for arguing that "demand is created by supply"; (2) for assuming that "the economic system was always operating up to its full capacity"; and (3) because Say's framework, Keynes believes, "is clearly incompetent to tackle the problems of unemployment and of the trade cycle."

When Keynes argues that, for Say, "demand is created by supply," Keynes ([1936] 1973, 18) takes it to mean that "in some significant but not clearly defined sense that the whole of the costs of production must necessarily be spent in the aggregate . . . on purchasing the product." According to Clower and Howitt (1998, 175–76n5), a possible source for Keynes's wording and formulation of Say's Law is John Stuart Mill's ([1844] 1874, 73) assertion: "Nothing is more true than that it is produce which constitutes the market for produce, and that every increase in production, if distributed without miscalculation among all kinds of produce in the proportion which private interest would dictate, creates, or rather constitutes, its own demand."[4]

In fact, Say's formulation was quite different. Say ([1814] 2006, 250) explained that "a product is no sooner created than it opens, from that instant, an outlet for other products to the full extent of its own value." The main difference between Say's formulation and Keynes's interpretation resides in the fact that Say referred to the *potential* of demand, in the sense that such a good did not *necessarily* create the demand because it was not necessarily sold, whereas Keynes implies the *automatic* creation of the demand for commodities. In addition, Say ([1803] 2006, 690; [1814] 2006, 250) explained that, besides products, the demand could also be for money or for financial assets.[5]

[3] The term "*débouchés*" is generally translated into English as "markets" or "vents." A better term is "outlets," which was used first by Lalor (1881–84, 3:38–40) and later by Baumol (1977, 147). Throughout this essay, translations of Say's writings are ours, unless otherwise noted.

[4] It should be noted that in the previous page of the same text, Mill ([1844] 1874, 72) was perfectly explicit. He described a situation where "money . . . was in request, and all other commodities were in comparative disrepute," thereby characterizing an excess supply of goods (excess demand for money).

[5] This point allows us to clarify an all-too-common confusion between Walras's Law and Say's Law. Walras's Law is derived from aggregating individual budget constraints, which implies that the total value of demand is equal to the total value of supply. When an individual determines the quantity of goods to be supplied and demanded on the basis of a given vector of prices, the individual assumes that he will be able to acquire the goods at these prices: selling what he supplies and buying what he demands. In contrast, Say reasoned in terms of equality between the value of the individual's resources and the value of his expenditure, which is radically different (as discussed in the next section).

The second criticism from Keynes alleges, "Say was implicitly assuming that the economic system was always operating up to its full capacity." This claim is untrue. As one example, in his *Lettres à M. Malthus*, Say (1820, 101n1) refuted David Ricardo's claim that "there is always as much industry as capital employed, and that all saved capital is always employed." Drawing upon the experience of the 1813 recession in France, Say contended that many savings were not invested, not all capital was employed, and many workers were jobless.[6] Though Say believed that saving was the engine of capital accumulation, he argued that many people had little to no savings and that substantial short-term change in the capital stock was unlikely. Say (1820, 73–74) wrote:

> It appears therefore . . . that one ought not, after Adam Smith, to preach parsimony. . . . In the first place, it is to be observed that most accumulations are necessarily slow. Everyone, whatever the income level, has to live before one can save; and what I here call *living*, is, in general, so much the more expensive as the individual is richer. In most cases and professions, the support of a family and its establishment in life exhaust the whole income, and often the capital besides; and when there are some yearly savings, they almost always represent a very small proportion of the capital actually employed . . . at any rate, only very great fortunes can allow great savings; and very great fortunes are rare in all countries. Therefore, capital can never augment with a rapidity capable of producing disruptions in industry.

Finally, Keynes concludes that Say's framework "is clearly incompetent to tackle the problems of unemployment and of the trade cycle." The concept of "unemployment" is not front and center in the writings of Say and other economists in the early nineteenth century, but Say was clearly aware that some individuals could not find employment and that the extent of flexibility of real wages could be part of the reason. For example, in a passage from *Lettres à M. Malthus*, Say (1820, 100–101n1) explained that a worker's labor services could not be hired because the subsistence wage was too high.[7] However, Say did not argue that money wages displayed downward rigidity, which is a distinctive characteristic of Keynes's framework (as noted by Modigliani 1944).

While Say did not analyze business cycles in the twentieth-century sense, he did analyze economic "crises." In the first four editions of *Traité*, like some other classical

[6] In *Traité*, Say ([1814] 2006, 260n1) said that "in 1813, manufacturing was in such a state of suffering, any type of industrial enterprise was so risky or not remunerative enough that capital could not be employed with acceptable security . . . the low rate of interest, which is ordinarily a sign of prosperity, was a sign of distress."

[7] Say (1820, 100–101n1) wrote: "The laborer can only sustain his work so long as his work pays for his subsistence; and when his subsistence is too costly, it no longer makes sense for any employer to hire him. It may then be said, in the language of political economy, that the laborer no longer *supplies* his productive services, even though he is eager to be employed; but this [labor] supply is not acceptable on the only lasting conditions on which it can be made" (emphasis in the original).

economists, such as Ricardo and Robert Torrens, Say held that crises were caused by a disproportion between supply and demand. Say later developed a theory of crises rooted in money and banking. In Say's discussion of the 1825 crisis in England, for example, he explained that banks discounted too many bills of exchange and overissued banknotes. Their clients panicked and tried to redeem their assets in cash, forcing banks to interrupt their discounting operations. Merchants could no longer obtain finance, businesses went bankrupt, and the situation escalated into a general crisis. Though Say did not offer a clear explanation of the recovery process, he nonetheless offered a penetrating interpretation of the causes of economic crises. Interestingly, Say's account suggests that he developed a theory of financial accelerators for the business cycle more than 150 years before the seminal work of Bernanke (1981, 1983) and others (for example, Bernanke, Gertler, and Gilchrist 1996) on the Great Depression.[8]

What Is Say's Law of Outlets?

In the nineteenth century, the common message of the defenders of what has come to be known as Say's Law was that production was the source of demand. However, that relationship need not imply that supply was necessarily equal to demand, nor that demand deficiency could not cause crises.[9]

Say ([1828–29] 2010, 349) defined outlets as "the means of trading products that [producers] have created for those that they need." Say's argument was that an increase in the value of production led to higher income and expenditures, which then generated greater outlets. Thus, Say (1820, 4–5) wrote, "as each of us can only purchase the products of others with his own products; as the value we can buy is equal to the value we can produce, the more men can produce, the more they will purchase." Say ([1819] 2006, 260) added, however, that in a stagnating or declining economy, "all the demands are on the decline; the value of the products is not equal to the costs of production." In other words, "while the demands are on the decline, there are always more goods [that are] supplied than goods [that are] sold" (Say [1814] 2006, 260n). Say ([1826] 2006, 1105) argued that "a buyer manifests himself in an effective manner only when there is money to buy; and he can obtain money only through the products that he created or those that were created for him; it then follows that it is production that stimulates outlets." Essentially, Say's law of outlets meant that selling goods increased one's holdings of money, which potentially, but not necessarily, allowed the purchase of other goods with the proceeds of the sale.

Say described a sequential model with consecutive transactions. An individual started off by selling productive services (which included labor) or goods that the

[8] In Say's discussion of the 1825 crisis in England, endogenous developments in banking and in credit markets seem to cause real and nominal shocks to the economy, not merely amplifying them as suggested in the modern literature.

[9] For a contrary interpretation, see Kates (1998, 2003, 2015).

person had produced. The individual then used the proceeds of the sale to purchase other products. In between the two transactions, money holdings for that individual could be greater or less than initial holdings. Say's reasoning can be summarized as follows: the total value of resources (initial money holdings for the individual plus the proceeds of the sale) was equal to the total value of expenditure (which included an individual's final money holdings). It should be noted that when Say used the term "product," he did not mean a physical quantity of material output, but rather an exchange value or market price (*prix courant*). In Say's mind, an unsold good or a good sold for less than its production cost did not constitute a "product." If some products were oversupplied, they would be sold for less than their production cost. The purchasing power of the producer would be reduced, and this loss would eventually affect expenditures by that producer. In his own words, "a product that does not reimburse its production costs, that is, a product whose monetary value does not cover profits and wages indispensable to satisfy [all] the needs . . . of consumers, is not a product, it is the inert result of a useless effort, at least so long as its monetary value remains below its production costs" (Say 1824, 28n1).

In Say's view, if some goods did not sell, it was because "many people bought less because they earned less" (Say [1814] 2006, 253). Demand was constrained by the amount of successful sales (Clower and Leijonhufvud [1973] 1981; Jonsson 1999; Béraud and Numa 2018a). Say thus recognized that the failure to produce (or the failure of factor owners to sell their services) must affect the demand for products, because that demand was financed out of earned income. One can contend that a glut could occur only in the short run, but it would still be a *general* glut (Hollander 2005a, 214–19; Hollander 2005b, 384). In *Cours*, Say ([1828–29] 2010, 196) reiterated that a general demand deficiency was possible and could cause a crisis and generate unemployment:

> In every country where manufacturing is very developed, there are moments where business is slow, and where the entire working class is suffering. This misfortune is not caused by the use of machinery, but by the nature of the manufactured products which are generally subject to multiple demand changes. These vicissitudes occur regardless of the methods used to make products, and they are even less dire in machine-intensive industries; because in industries where everything is done with manpower, if jobs are lacking, many men are deprived of food, whereas when a machine is idle, its owner loses only the interest on capital that it represents.

Clearly, Say does not overlook the possibility of demand changes leading to unemployment. At the end of *Cours*, after stressing the importance of saving and reproductive consumption (what modern economists would call "investment"), Say ([1828–29] 2010, 1231–32) opined that the purpose of economic life was to consume, but not necessarily today; if individuals wanted to enjoy more goods and services tomorrow, they had to save. Thus, in Say's mind, economic policy should aim to stimulate production rather than demand to ensure prosperity. Indeed, "consumption is not a cause: it is an effect. One must buy in order to consume; yet one can buy only what has been produced. The quantity of products demanded is therefore determined by the quantity of products created? Undoubtedly so" (Say [1803] 2006, 688).

Money, Saving, and Hoarding

A number of commentators have alleged that Say conceived of money only as a medium of exchange. For example, Schumpeter (1954, 590) erroneously claims that "Say . . . did not consider the problem of hoarding" and concludes "that Say, neglected the store-of-value function of money." In reality, Say did study the function of store of value, and he did discuss hoarding in several instances in his writings (Numa forthcoming [b]). Moreover, Say clearly understood that changes in the value of money affected the real economy.

For Say, if individuals wish to increase their money holdings, money demand might exceed money supply. In this case, one of two scenarios will take place. If money is lacking, monetary substitutes such as bills of exchange, promissory notes, or other credit instruments will be used (Say [1814] 2006, 248). If the excess demand for money persists, the value of money will rise. In an open economy in which money is convertible into gold at a fixed rate, an inflow of gold will lead to an increase in the quantity of money (as David Hume described).

Motives for Holding Money

Keynes's message in the *General Theory* is that production does not always generate a demand for other products, because the existence of money can create a gap between savings and investment. In Keynes's terminology, there are three motives behind the individual desire to hold money: the "transaction-motive," the "precautionary-motive," and the "speculative-motive" (Keynes [1936] 1973, 170). The first two motives depend on the level of income, while the third motive depends on the interest rate. Hoarding is based on the speculative motive (170, 196–99). In Keynes's theory the interest rate is "the reward for parting with liquidity"—that is, it is "the 'price' which equilibrates the desire to hold wealth in the form of cash with the available quantity of cash" (167).

Say identified three quite similar motives behind the desire to hold money. First, Say ([1828–29] 2010, 400) described an income-elastic demand for money for transaction purposes: "What quantity of money will I need? The more sales and purchases I will have to carry out, the more money I will need. The manufacturer who needs to sell and purchase for an amount of five thousand francs every year, will use, in the course of a year, much more money than the porter who only receives in wage and consumes a thousand francs in the same time period." Second, Say referred to a money demand to deal with unforeseen contingencies: "there are some types of occupation and consumption that always require to keep . . . a certain sum to deal with unforeseen expenses" (401). These first two motives show that, for Say, the demand for money depended upon the level of income. For the third motive, Say wrote that "as one loses interest in holding money, I assume that no one holds more [money] than one expects to use" (401), and added in a footnote that "the money used . . . to cover expenses inherent to the movement of business, is part of the capital of the firm; and the portion of money that remains idle . . . is unproductive capital." Say's reasoning displays an interest-elastic motive for money demand which indicates that the interest rate is a reward for parting with cash, in line with Keynes's approach.

Savings and Hoarding

In *A Treatise on Money*, Keynes ([1930] 1971, 2:127) defines hoarding as the holding of money, including bank deposits. A classical economist would contend that the banking system theoretically serves as an intermediary between lenders and borrowers. However, during downturns—generally characterized by greater uncertainty—banks do not necessarily reallocate the funds collected. Say stated that during the 1813 recession in France, "the Bank of France alone [had] 223 million in cash in his vaults, an amount which is worth more than twice the sum of its bills in circulation, and six times greater than what prudence would recommend to face potential demand of redemption/reimbursement" (Say 1820, 102n1).

In the *General Theory*, hoarding is expanded into the concept of liquidity preference. Keynes ([1937] 1973, 116) is known for emphasizing money's store-of-value function in situations of uncertainty by stating that "our desire to hold money as a store of wealth is a barometer of the degree of our distrust of our own calculations and conventions concerning the future." In a similar spirit, Say ([1828–29] 2010, 149) stated that "the lack of security and confidence often leads owners of capital funds to refrain from investing for fear of compromising them. They prefer to lose interest instead of risking the principal." Say explained that owners of idle funds factored risk and return in their decision-making. Hoarding could thus involve large amounts. For Say, the interest rate not only affected the decision to hold bills of exchange and promissory notes but also determined how long individuals held such assets.

Like Keynes, Say ([1803] 2006, 204) emphasized the role of uncertainty in decisions to hold cash balances. If the current market environment was too uncertain and/or risky or if some profits were expected in the future, it made sense to hold on to cash balances:

> When industry was at an early stage, an unprofitable capital was almost nothing but a treasure kept in a coffer or buried underground . . . in case of a need; significant or not, this treasure did not generate more or less profit, because it gave none; it was nothing but some sort of precaution. But when the treasure generated a profit commensurate with its mass then people became incentivized to make it grow. And this was not motivated by a loose interest, based on a precautionary motive, but based on a true interest, that could be felt anytime, because the profit generated by the capital could be spent and allow new uses without being destroyed.

The passage suggests that hoarding could involve large amounts in underdeveloped countries, as individuals kept money idle because of a lack of investment opportunities. In that case, it was rational to wait until the expected profit was greater than the amount of cash hoarded. In *Cours*, Say paid greater attention to manufacturing instability and depressions. He reiterated that hoarding made perfect sense, but in this case hoarding arose as a result of a lack of information: “If some individuals hoard, we can consider that they strive to keep a treasure in reserve as a result of a need; and it can be argued that these individuals usually feel the need to keep with them a certain amount of [money] that better-advised individuals can employ to a better use” (Say [1828–29] 2010, 401). Hoarding was, in short, an integral part of Say’s economic system.

Of course, these similarities do not mean that Keynes and Say held identical views on the role of money in economic downturns. One difference is that Say argued that the demand for money for precautionary motive rose during depressions, indicating that hoarding was a *symptom* rather than a cause of depressions (although this view is also held by some Keynesian economists, such as Rowe 2016). Say was convinced that any “treasure” of hoarded unprofitable would get spent eventually, perhaps over the very long run (more than one generation). This also explains why, despite acknowledging that hoarding rose during depressions, Say never prescribed any remedy to curb hoarding, unlike Keynes ([1936] 1973, 353–56), who espoused Silvio Gesell’s idea of “stamped” money, a sort of tax to impose a direct cost on money holders for refusing to part with cash (Darity 1995, 27).

Another difference is that Keynes seems to view the interest rate as a pure monetary phenomenon, while Say believed that the interest rate was the price determined by the supply of and the demand for loanable funds (Say [1828–29] 2010, 401n1, 1229). In fact, Say ([1828–29] 2010, 479) embraced Thomas Tooke’s (1826, 22–24) views that money temporarily affected the interest rate. Additional money in the economy brought more lending, which resulted in a larger amount of capital funds that pushed the rate of interest down and subsequently lowered production costs. Prices then rose because of the abundance of money, but this effect occurred after the decline of the interest rate. Given that producers purchased their inputs before prices climbed, they profited from a low cost of borrowing. As they sold their products when prices went up, they ended up making large profits.

Money Is Not Neutral

In the fifth edition of *Traité* (Say [1826] 2006, 505) and in *Cours* (Say [1828–29] 2010, 479), Say studied the effects of a growing quantity of money. He noted that more money stimulated all sales and thereby boosted the demand for goods for two reasons: final prices outran production costs, giving larger profits to producers, and inflation expectations led consumers to spend money more rapidly.

Intriguingly, Say’s analysis was very similar to Keynes’s discussion in *A Treatise on Money*, in which profit inflation occurred when prices were outrunning costs, leaving a large and growing margin for profit (Keynes [1930] 1971, 2:137). Moreover, Keynes referred to gently rising prices just as did Say, who alluded to a gradual and moderate price increase. Say's claim was based on the secular record. He acknowledged that the greater quantity of banknotes and inconvertible paper money, respectively, during the early stages of John Law's scheme and the early days of *assignats* had expansionary effects on the French economy.[10] Say ([1828–29] 2010, 479) concluded that "in spite of the principles that teach us that money plays only the role of a simple intermediary, and that products can ultimately be purchased only with products, more abundant money fosters all sales and the reproduction of new values."

Government Intervention and Public Works

In the conclusion of the *General Theory*, Keynes ([1936] 1973, 379) makes it clear that his argument is for government "to succeed in establishing an aggregate volume of output corresponding to full employment as nearly as is practicable." But if the overall aggregate volume of output is sufficient, then market forces should be allowed to function to promote efficiency and freedom. Keynes writes:

> To put the point concretely, I see no reason to suppose that the existing system seriously misemploys the factors of production which are in use. . . . When 9,000,000 men are employed out of 10,000,000 willing and able to work, there is no evidence that the labour of these 9,000,000 men is misdirected. The complaint against the present system is not that these 9,000,000 men ought to be employed on different tasks, but that tasks should be available for the remaining 1,000,000 men. It is in determining the volume, not the direction, of actual employment that the existing system has broken down.

This statement does not sound very different from Say's advocacy for a hands-off approach to economic affairs, in his case mainly directed toward Napoléon Bonaparte's tyrannical regime.

In the *General Theory*, Keynes ([1936] 1973, 378) argues "that a somewhat comprehensive socialisation of investment will prove the only means of securing an approximation to full employment; though this need not exclude all manner of compromises and of devices by which public authority will co-operate with private initiative." In other writing, Keynes spelled out in more detail what he had in mind by public investment. In a 1929 election pamphlet that Keynes authored with Hubert Douglas Henderson, their stimulus plan included public spending in transportation, housing, energy, and telecommunications. The rationale for government intervention is articulated as follows (Keynes and Henderson [1929] 1972, 113; emphasis in the original):

[10] *Assignats* were initially issued as treasury bonds in 1790 but then circulated as inconvertible paper money from 1791 to 1796.

> Why must the Government play a part itself? Why is it not enough to offer facilities and encouragement to private enterprise? . . . Whether we like it or not, *it is a fact* that the rate of capital development in the transport system, the public utilities and the housing of this country largely depends on the policy of the Treasury and the Government of the day. . . . The choice between a well-equipped, up-to-date, go-ahead and efficient national plant depends on the mood and policy of the Government. Thus it is not a question of choosing between private and public enterprise in these matters. The choice has been already made. In many directions—though not in all—it is a question of the State putting its hand to the job or of its not being done at all. Roads, afforestation, reclamation and drainage, electrification, slum clearance and town planning, the development of canals, docks and harbours; these are the things which need to absorb large sums of capital to-day, and in every case the initiative necessarily lies with a public authority.

In short, Keynes ([1933] 1982, 158) argued that the purpose of government intervention was to step in for failing private initiative in order to "break the vicious circle."

For his part, Say often criticized government intervention in private affairs in general. However, from the first edition of *Traité*, Say ([1803] 2006, 330) recognized that "there are circumstances that can modify this generally true proposition that everyone is the best judge of how to use his industry and capital." In these situations, self-interest became ineffective and socially undesirable, and public intervention was required (Numa forthcoming [a]). As one example, Say ([1803] 2006, 329–30; [1814] 2006, 63) argued that the government could grant temporary protection for infant industries facing international competition.

Say also pushed for government intervention in the form of public works as a remedy for unemployment resulting from the introduction of machinery (Baumol 1997). He suggested industrial policy so the government could confine the use of new machines in regions where labor was scarce. Say ([1803] 2006, 136–37) also suggested creating companies with public funds in order to give jobs to unemployed individuals, and thereby to jump-start the economy:

> Note that a clever administration can always find ways to alleviate this temporary and local evil. In the early stages, it can restrict the use of a new machine to certain areas where labor is scarce and demanded by other sectors of industry. It can provide in advance unemployed individuals with some employment, by forming companies of public utility with its own funds, such as those in charge of a canal, a road, a major building.[11]

[11] In the fourth edition of *Traité*, Say ([1819] 2006, 137n1) invoked a "benevolent administration" instead of a "clever administration."

Say ([1814] 2006, 385) was perfectly clear in his support of stimulating the private sector with the underpinning of a public works program, writing that "the government is a bad producer . . . yet it could powerfully stimulate private production with well-designed public establishments, properly executed and well-maintained, and especially with roads, bridges, canals and ports." In Say's thinking, public infrastructure boosted productivity and spurred economic growth. Say ([1819] 2006, 171; [1826] 2006, 167) added that "this is the reason why roads, canals, bridges . . . [and] everything that facilitates domestic communications, enhance the wealth of a country" (see also Say [1803] 2006, 388). In general, Say criticized public debt because he feared that the funds would be used for wasteful expenditures and unproductive consumption, such as funding wars and military purchases. However, he welcomed public debt for "the construction of bridges, the construction or maintenance of roads and canals, and all public infrastructure indirectly productive" (Say [1803] 2006, 766; see also Say [1828–29] 2010, 1007–8).

On the issue of how demand-side policies such as public works can be effective because they enhance productive capacities for the economy as a whole, Littleboy (2003, 165) has argued that Keynes's and Say's "systems dovetail . . . the policy implications overlap." He points out that when talking about how government-funded infrastructure encourages private investment, the demand-side versus supply-side debate "loses its energy."

Conclusion

In our view, there are enough similarities in their analyses to call into question the idea that the views of John Maynard Keynes were antithetical to those of Jean-Baptiste Say. Indeed, Keynes could have readily agreed with the Frenchman on several issues, such as the possibility of aggregate-demand deficiency, the role of money in the economy, and government intervention through public works. Of course, Keynes and Say were also writing a century apart, with meaningful differences in their approaches. While poverty was the main issue for economists in the early nineteenth century, unemployment was the main concern in the 1930s when Keynes was writing his *General Theory*. Keynes was building a macroeconomic model of an economy with less than full employment under conditions of money-wage stickiness, while the idea of undertaking such a model would not have been within the worldview of early nineteenth-century economists.

Since the publication of Keynes's *General Theory*, generations of economists have been told that Keynes and Say were polar opposites, and that Keynes was the ultimate nemesis for Say. This perspective was from the start built on Keynes's misinterpretation of Say's views. Our investigation has brought to the fore a much more complex and overlapping set of relationships between the theories of these two giants.

■ *We thank Gordon Hanson, Timothy Taylor, and Heidi Williams for their helpful comments. Some parts of the present essay draw upon Béraud and Numa (2018b).*

References

Baumol, William J. 1977. "Say's (at Least) Eight Laws, or What Say and James Mill May Really Have Meant." *Economica* 44(174): 145–61.

Baumol, William J. 1997. "J.-B. Say on Unemployment and Public Works." *Eastern Economic Journal* 23(2): 219–30.

Baumol, William J. 1999. "Retrospectives: Say's Law." *Journal of Economic Perspectives* 13(1): 195–204.

Béraud, Alain, and Guy Numa. 2018a. "Beyond Say's Law: The Significance of J.-B. Say's Monetary Views." *Journal of the History of Economic Thought* 40(2): 217–41.

Béraud, Alain, and Guy Numa. 2018b. "Keynes, J.-B. Say, J. S. Mill, and Say's Law: A Note on Kates, Grieve, and Ahiakpor." *Journal of the History of Economic Thought* 40(2): 285–89.

Bernanke, Ben S. 1981. "Bankruptcy, Liquidity, and Recession." *American Economic Review* 71(2): 155–59.

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

Bernanke, Ben S., Mark Gertler, and Simon Gilchrist. 1996. "The Financial Accelerator and the Flight to Quality." *Review of Economics and Statistics* 78(1): 1–15.

Blanc, Emmanuel, and André Tiran. 2003. "Introduction générale." In *Œuvres morales et politiques*, 9–42. Vol. 5 of *Œuvres Complètes de Jean-Baptiste Say*. Paris: Economica.

Clower, Robert Wayne. 2004. "Trashing J. B. Say: The Story of a Mare's Nest." In *Macroeconomic Theory and Economic Policy: Essays in Honour of Jean-Paul Fitoussi*, edited by K. Vela Vilupillai, 88–97. London: Routledge.

Clower, Robert, and Peter Howitt. 1998. "Keynes and the Classics: An End of Century View." In *Keynes and the Classics Reconsidered*, edited by James C. W. Ahiakpor, 163–78. Boston: Kluwer.

Clower, Robert, and Axel Leijonhufvud. (1973) 1981. "Say's Principle: What It Means and Doesn't Mean." *Intermountain Economic Review* 4(2): 1–16. Reprinted in *Information and Coordination. Essays in Macroeconomic Theory*, edited by Axel Leijonhufvud, 79–101. Oxford: Oxford University Press.

Darity, William Jr. 1995. "Keynes' Political Philosophy: The Gesell Connection." *Eastern Economic Journal* 21(1): 27–41.

Forget, Evelyn L. 1999. *The Social Economics of Jean-Baptiste Say: Markets and Virtue*. London: Routledge.

Hollander, Samuel. 2005a. *Jean-Baptiste Say and the Classical Canon in Economics. The British Connection in French Classicism*. London: Routledge.

Hollander, Samuel. 2005b. "Two Hundred Years of Say's Law: Essays on Economic Theory's Most Controversial Principle, edited by Steven Kates." *History of Political Economy* 37(2): 382–85.

Jacoud, Gilles. 2013. *Money and Banking in Jean-Baptiste Say's Economic Thought*. Abingdon, UK: Routledge.

Jonsson, Petur O. 1997. "On Gluts, Effective Demand, and the True Meaning of Say's Law." *Eastern Economic Journal* 23(2): 203–18.

Jonsson, Petur O. 1999. "'Say's Law and the Keynesian Revolution: How Macroeconomics Lost Its Way' by Steven Kates." *Southern Economic Journal* 65(4): 967–70.

Kaplan, Andreas. 2014. "European Management and European Business Schools: Insights from the History of Business Schools." *European Management Journal* 32(4): 529–34.

Kates, Steven. 1998. *Say's Law and the Keynesian Revolution: How Macroeconomic Theory Lost Its Way*. Northampton, MA: Edward Elgar.

Kates, Steven, ed. 2003. *Two Hundred Years of Say's Law*. Northampton, MA: Edward Elgar.

Kates, Steven. 2015. "Mill's Fourth Fundamental Proposition on Capital: A Paradox Explained." *Journal of the History of Economic Thought* 37(1): 39–56.

Keynes, John Maynard. (1930) 1971. *A Treatise on Money*. London: Macmillan. Reprinted in *A Treatise on Money: The Pure Theory of Money* and *A Treatise on Money: The Applied Theory of Money*. Vols. 5 and 6 of *The Collected Writings of John Maynard Keynes*. London: Macmillan, for the Royal Economic Society.

Keynes, John Maynard. (1933) 1982. "A Programme for Unemployment," in *Activities 1931–1939: World Crises and Policies in Britain and America*, 154–61. Vol. 21 of *The Collected Writings of John Maynard Keynes*. London: Macmillan, for the Royal Economic Society.

Keynes, John Maynard. (1936) 1973. *The General Theory of Employment, Interest and Money*. London: Macmillan. Reprinted in *The General Theory*. Vol. 7 of *The Collected Writings of John Maynard Keynes*. London: Macmillan, for the Royal Economic Society.

Keynes, John Maynard. (1937) 1973. "The General Theory of Employment." *Quarterly Journal of Economics* 51(2): 209–23. Reprinted in *The General Theory and After: Part II. Defence and Development*, 109–23. Vol. 14 of *The Collected Writings of John Maynard Keynes*. London: Macmillan, for the

Royal Economic Society.

Keynes, John Maynard. (1939) 1973. "Preface to the French Edition." In *The General Theory of Employment, Interest and Money*. London: Macmillan. Reprinted in *The General Theory*, pp. xxxi–xxxv. Vol. 7 of *The Collected Writings of John Maynard Keynes*. London: Macmillan, for the Royal Economic Society.

Keynes, John Maynard, and Hubert Douglas Henderson. (1929) 1972. *Can Lloyd George Do It?* London: The Nation and Athenæum. Reprinted in *Essays in Persuasion*, 86–125. Vol. 9 of *The Collected Writings of John Maynard Keynes*. London: Macmillan, for the Royal Economic Society.

Lalor, John Joseph, ed. 1881–84. *Cyclopædia of Political Science, Political Economy, and of the Political History of the United States*. 3 vols. New York: Charles Merrill and Co.

Littleboy, Bruce. 2003. "Say's Lore." In Kates 2003, 154–67.

Lutfalla, Michel. 1991. "Jean-Baptiste Say, 1767–1832, le fondateur." In *L'économie politique en France au XIXe siècle*, edited by Yves Breton and Michel Lutfalla, 13–31. Paris: Economica.

Mill, John Stuart. (1844) 1874. *Essays on Some Unsettled Questions of Political Economy*, 2nd ed. London: Longmans.

Modigliani, Franco. 1944. "Liquidity Preference and the Theory of Interest and Money." *Econometrica* 12(1): 45–88.

Moggridge, Donald Edward. 1992. *Maynard Keynes: An Economist's Biography*. London: Routledge.

Numa, Guy. Forthcoming (a). "Jean-Baptiste Say on Free Trade." *History of Political Economy*. http://ssrn.com/abstract=3395810.

Numa, Guy. Forthcoming (b). "Money as a Store of Value: J.-B. Say on Hoarding and Idle Balances." *History of Political Economy*. http://ssrn.com/abstract=3395821.

Palmer, Robert Roswell. 1997. *J.-B. Say: An Economist in Troubled Times*. Princeton, NJ: Princeton University Press.

Patinkin, Don. 1976. *Keynes's Monetary Thought: A Study of Its Development*. Durham, NC: Duke University Press.

Potier, Jean-Pierre. 2010. "Introduction." In Say (1828–29) 2010, pp. ix–lvi.

Rowe, Nicholas. 2016. "Keynesian Parables of Thrift and Hoarding." *Review of Keynesian Economics* 4(1): 50–55.

Say, Jean-Baptiste. (1803, 1814, 1817, 1819, 1826, 1841) 2006. *Traité d'économie politique, ou simple exposition de la manière dont se forment, se distribuent et se consomment les richesses*. Variorum edition in *Traité d'économie politique*. Vol. 1 (2 bks.) of *Œuvres Complètes de Jean-Baptiste Say*. Paris: Economica.

Say, Jean-Baptiste. 1820. *Lettres à M. Malthus sur différents sujets d'économie politique notamment sur les causes de la stagnation générale du commerce*. Paris: Londres Bossange.

Say, Jean-Baptiste. 1824. "Balance des productions avec les consommations." *Revue Encyclopédique* 23(July): 19–31.

Say, Jean-Baptiste. (1828–29) 2010. *Cours complet d'économie politique pratique*. Paris: Rapilly. Variorum edition in *Cours complet d'économie politique pratique*. Vol. 2 (2 bks.) of *Œuvres complètes de Jean-Baptiste Say*. Paris: Economica.

Schoorl, Evert. 2013. *Jean-Baptiste Say: Revolutionary, Entrepreneur, Economist*. Abingdon, UK: Routledge.

Schumpeter, Joseph A. 1954. *History of Economic Analysis*. New York: Oxford University Press.

Skidelsky, Robert. 2005. *John Maynard Keynes, 1883–1946: Economist, Philosopher, Statesman*. New York: Penguin Books.

Steiner, Philippe. 1990. "L'économie politique pratique contre les systèmes: quelques remarques sur la méthode de J.-B. Say." *Revue d'économie politique* 100(5): 664–87.

Summers, Lawrence. 2014. "The Path to Full Employment: Making Jobs a National Priority." Keynote speech given at the Center on Budget and Policy Priorities, Washington, DC, April 2, 2014. https://www.youtube.com/watch?v=oMsxPN5bCCA.

Tooke, Thomas. 1826. *Considerations on the State of the Currency*. London: John Murray.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 243–248

Some *Journal of Economic Perspectives* Articles Recommended for Classroom Use

Timothy Taylor

In the first half of 2018, the editors of the *Journal of Economic Perspectives* sent out several invitations, in the back pages of the journal and via email blasts from the American Economic Association, for faculty to send us examples of JEP articles that they had found useful for their syllabus or other classroom uses.

For some JEP readers, the request raised concerns. One wrote: "I need to ask about this information collection you're engaged in. I am inferring it is because the journal does not look as good based on traditional performance metrics and you're trying to justify its value to the AEA publication board. Is that correct?"

Fortunately for peace of mind in our editorial offices, the JEP does just fine on traditional metrics of journal performance like citation counts. For example, according to the InCites Journal Citation Reports published by Clarivate Analytics, the JEP ranked between third and fifth during the five most recent available years, from 2014 to 2018, among all academic journals of economics in "Journal Impact Factor," which is a measure of how often articles published in the past two years have been cited in the academic literature the following year, divided by the total number of articles (https://jcr.clarivate.com/; accessed July 2, 2019; log-in required). Scopus (from Elsevier) calculates a CiteScore, which is "calculated from all citations recorded in Scopus in one year to content published in the last three years, divided by the number of items published." By this measure, the JEP ranked fourth in 2018 in the broad category of Economics, Econometrics, and Finance (https://www.scopus.com/sources; accessed July 2, 2019). Google Scholar calculates the "h5-index," the

■ *Timothy Taylor is Managing Editor,* Journal of Economic Perspectives, *based at Macalester College, Saint Paul, Minnesota. His email address is taylort@macalester.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.243 doi=10.1257/jep.33.3.243

largest number h such that h articles published in 2013–2017 have at least h citations each. The JEP ranks seventh among all economics journals by this measure (https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=bus_economics; accessed July 2, 2019).

But while citations to JEP articles are very welcome, the journal also aspires to fulfill some broader functions: as it says in the "Statement of Purpose" on the second page of each issue, it "attempts to fill a gap between the general interest press and most other academic economics journals." Indeed, this broader mission was part of the reason why the American Economic Association decided back in 2011 to make the JEP freely available online. In 2018, there were about 1.5 million downloads of JEP articles from the AEA website. One reason for the choice to make JEP freely available was to make it easier for faculty to assign JEP articles to students.

When we invited faculty members to send JEP articles that had proved useful in their classrooms, with a focus on their undergraduate courses, we worried about receiving only a handful of replies and hoped to receive at least a few dozen. We ended up receiving 250 responses, many recommending multiple JEP articles for classroom use and some including syllabuses and cover letters with additional comments.

The Categories

On the JEP website, we have created a landing page (https://www.aeaweb.org/journals/jep/classroom) that organizes the recommended articles into 33 categories. (This resource can also be accessed from the article page on the JEP website.) Many of them refer to specific courses, while others may be more appropriately thought of as subject headings. Of course, this classification involved a number of judgement calls. The list of categories appears in Table 1. If you visit the link above and click on any of the categories, you will see a list of papers from the JEP that were recommended by faculty members for classroom use for that category, listed in reverse date order. Each article listed includes a hyperlink to its article page on the JEP website.

A few thoughts about how this exercise was carried out, along with its strengths and limitations, seem appropriate.

First, we make no pretense of suggesting or providing a complete syllabus for any specific course. We offer only the milder hope that these recommendations from peers might suggest some additional readings for your students.

Second, there were obvious issues when categorizing papers and avoiding an undue amount of duplication. Closely related classes can have different names. Certain papers were recommended for multiple classes. As one example, many of the same papers were recommended for classes in Intermediate Macroeconomics, Money and Banking, and Financial Markets. Many of the papers listed under Econometrics turned up in a variety of other classes as well. Papers about China, for example, could be listed under Development, or under specific subjects such as Labor Economics, Environment/Energy, or Intermediate Macroeconomics. We did

Table 1
Categories for JEP Articles in the Classroom
(for links, go to https://www.aeaweb.org/journals/jep/classroom)

Principles or Introductory Course
Intermediate Macroeconomics
Financial Markets
Econometrics
Labor Economics
Education
Environment/Energy
Game Theory
Industrial Organization
Household Economics
Immigration and Emigration
Urban Economics
Europe: Topics Course
Soviet and Post-Soviet: Topics Course
Latin America: Topics Course
Public Policy
Economics Profession

Intermediate Microeconomics
Money and Banking
International
Experimental Methods
Health Economics
Public Finance
Behavioral Economics
Social Norms and Networks
Law and Economics
Development
Economic History
Sports Economics
China: Topics Course
Japan: Topics Course
Middle East: Topics Course
Political Economy

Notes: By following the link above, you will find a landing page listing these categories. Click on any of the categories at the landing page, and you will find a list of JEP articles in that area that have been recommended by surveyed faculty members.

not try to eliminate all duplication, and some papers appear in two or even three categories. But if we had not taken steps to limit the extent of duplication, or to create some categories like China: Topics Course, the number of entries for some of these categories would have been at least twice as large.

Third, there were many cases where a faculty member referred to a symposium, but not to individual papers. In those cases, we just listed all the papers from the symposium. However, if only one paper from a symposium was mentioned, we listed only that single paper.

Finally, our requests for suggested articles went out between February and May 2018. Thus, while the list at the website does include a few articles from late in 2017 or early in 2018, it tends to be focused on the pre-2018 period. In particular, the listings for the International class do not include the "Symposium on Does the US Really Gain from Trade?" from the Spring 2018 issue, and the Behavioral Economics class listings do not include the "Symposium on Risk in Economics and Psychology" from the same issue. The course listings for Intermediate Macroeconomics do not include the "Symposium on Macroeconomics a Decade after the Great Recession" from Summer 2018. The course listings for Environment/Energy do not include the "Symposium on Climate Change" from Fall 2018, and the course listings for Public Finance do not include the "Symposium on the Tax Cuts and Jobs Act" from that issue, either. The same point can be made for the issues of 2019, as well. The purpose of this list was to pass along recommendations from faculty, not to compile a systematic list from back issues of the JEP, and we stuck to that mission.

How Frequently?

The compilation of recommended articles at the JEP website neither addresses the question of how frequently JEP articles appear on syllabuses nor reveals which articles appear most frequently. With regard to the first question, then-editor David Autor (2012) described a Google search of the websites of the top 100 US research universities, using the terms "Journal of Economic Perspectives" and "syllabus." He found an average of 43 JEP articles on syllabuses at each school in 2010. He also noted that this total list is likely an underestimate, because many syllabuses are not readily available online. In addition, his search was done a year before the JEP became freely available online in 2011.

For a perspective on which JEP articles are currently most likely to appear on course syllabuses, we turned to the Open Syllabus Project, managed by the American Assembly at Columbia University. When we went to the Open Syllabus Project (http://explorer.opensyllabusproject.org/; accessed July 2, 2019) and used "Journal of Economic Perspectives" as the search term, we found 799 JEP articles listed. The top 30 appear in Table 2. The list, and in particular the counts of how many syllabuses, should be treated only as suggestive. The website notes: "At present, we have around 1.1 million syllabi, drawing predominantly from the past decade of teaching in the US. We think the total number of US, UK, Canadian, and Australian syllabi for the past 15 years is in the range of 80–100 million." It's easy to imagine that some course syllabuses are being counted in several different years, while 90 percent or more of syllabuses are not being counted at all.

However, it's interesting to note some prominent examples of JEP articles being used in noneconomics courses. For example, the third entry in Table 2 is "Legislative Organization" by Keith Krehbiel, which appeared in the Winter 2004 issue as part of a four-paper "Symposium on Political Economy." However, none of the 250 economists who offered suggestions mentioned the 2004 article by Krehbiel. We suspect that it is being used in political science classes.

Lessons for the Editors

The feedback and suggestions offered some lessons for us at the JEP, as well. Many respondents just included the JEP papers that appear on their reading lists, but here are some of the other uses mentioned in correspondence.

First, JEP articles are frequently used as a basis for structured student writing or discussion assignments. A number of faculty members described doing this, or sent along their instructions for such assignments. A typical approach was to ask students to summarize key arguments, theories, and evidence—sometimes in writing, sometimes verbally. Sometimes students were asked to contrast the arguments in several JEP papers.

Second, a substantial number of faculty members used JEP papers in junior- or senior-level seminar-style classes. We did not include "Research Seminar" as

Table 2
JEP Articles Most Likely to Appear: Open Syllabus Project

Rank	*Count*	*Article*
1	148	"Divergence, Big Time" by Lant Pritchett (Summer 1997)
2	130	"Are Your Wages Set in Beijing?" by Richard B. Freeman (Summer 1995)
3	123	"Legislative Organization" by Keith Krehbiel (Winter 2004)
4	89	"The Origins of Endogenous Growth" by Paul M. Romer (Winter 1994)
5	89	"Capital Structure" by Stewart C. Myers (Spring 2001)
6	82	"The Contingent Valuation Debate: Why Economists Should Care" by Paul R. Portney (Fall 1994)
7	77	"The Nation in Depression" by Christina D. Romer (Spring 1993)
8	77	"Political Regimes and Economic Growth" by Adam Przeworski and Fernando Limongi (Summer 1993)
9	75	"Valuing the Environment through Contingent Valuation" by W. Michael Hanemann (Fall 1994)
10	68	"Medical Care Costs: How Much Welfare Loss?" by Joseph P. Newhouse (Summer 1992)
11	66	"Integration of Trade and Disintegration of Production in the Global Economy" by Robert C. Feenstra (Fall 1998)
12	62	"Government Failures in Development" by Anne O. Krueger (Summer 1990)
13	60	"The Global Capital Market: Benefactor or Menace?" by Maurice Obstfeld (Fall 1998)
14	59	"Are Cities Dying?" by Edward L. Glaeser (Spring 1998)
15	58	"Real Business Cycles: A New Keynesian Perspective" by N. Gregory Mankiw (Summer 1989)
16	58	"The Boundaries of Multinational Enterprises and the Theory of International Trade" by James R. Markusen (Spring 1995)
17	57	"How Costly Is Protectionism?" by Robert C. Feenstra (Summer 1992)
18	57	"Reflections on the Economics of Climate Change" by William D. Nordhaus (Fall 1993)
19	57	"Why Has Africa Grown Slowly?" by Paul Collier and Jan Willem Gunning (Summer 1999)
20	56	"Understanding Real Business Cycles" by Charles I. Plosser (Summer 1989)
21	55	"Evidence on Discrimination in Mortgage Lending" by Helen F. Ladd (Spring 1998)
22	54	"Collective Action and the Evolution of Social Norms" by Elinor Ostrom (Summer 2000)
23	53	"Does the 'New Economy' Measure Up to the Great Inventions of the Past?" by Robert J. Gordon (Fall 2000)
24	52	"Can Foreign Aid Buy Growth?" by William Easterly (Summer 2003)
25	52	"Auctions and Bidding: A Primer" by Paul Milgrom (Summer 1989)
26	50	"The Case for Randomized Field Trials in Economic and Policy Research" by Gary Burtless (Spring 1995)
27	50	"The Political Economy of Trade Policy" by Robert E. Baldwin (Fall 1989)
28	50	"On the Evolution of the World Income Distribution" by Charles I. Jones (Summer 1997)
29	48	"Contingent Valuation: Is Some Number Better Than No Number?" by Peter A. Diamond and Jerry A. Hausman (Fall 1994)
30	48	"The Worldwide Standard of Living since 1800" by Richard A. Easterlin (Winter 2000)

Source: Search for "Journal of Economic Perspectives" on the Open Syllabus Project Explorer (beta 0.4, http://explorer.opensyllabusproject.org/), performed on July 2, 2019.

a separate category on the master list (shown above in Table 1), because such a category would have included several hundred articles from all different fields and areas. Often, the purpose of JEP papers in such courses seemed to be to give students a set of readings that let them use the terminology and analysis they had learned in earlier classes and to help launch students into their own research projects.

Third, one unexpected finding was that a number of faculty members are using the appendices of certain JEP articles as a basis for quantitative classroom exercises. Sometimes this occurs in an econometrics class; sometimes in other courses. Students may start by replicating the results of a JEP paper. Then they may rerun the analysis in some different way: perhaps by downloading different or updated data, or by trying an alternative statistical specification. The lesson here for the editors is that we should pay greater attention to what we request in appendices. If the authors of JEP papers know that their appendices may be used as the basis for a classroom exercise, they can structure the material and provide an appropriate level of detail with that use in mind.

Fourth, it's worth having the JEP return to prominent topics perhaps every five years or so. Many faculty members mentioned that they value papers that are relatively up to date, and some mentioned that they had used certain JEP articles for a time but eventually stopped because the paper felt aged.

Fifth, numerous respondents offered suggestions for specific topics to be addressed, or to be addressed again because a previous symposium had become dated.

As a final lesson, many readers took the time to write notes offering some pleasingly positive feedback about the JEP and how they made use of it not only on reading lists but as background for lectures and to keep up with the field of economics as a whole. Of course, we recognize that those who took the time to write are the definition of a nonrandom sample, and we are thus prohibited by the social scientists' creed from drawing any generalized conclusions about the JEP from these responses. But we do very much appreciate the kind words.

■ *Thanks to Andra Boca for organizing the survey results and to Jocelyn Rice and the staff at the American Economic Association for their work on the web pages listing the 33 categories and the underlying JEP articles.*

References

Autor, David. 2012. *The Journal of Economic Perspectives* at 100 (Issues). *Journal of Economic Perspectives* 26(2): 3–18.

Journal of Economic Perspectives—Volume 33, Number 3—Summer 2019—Pages 249–256

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives,* Macalester College, 1600 Grand Ave., St. Paul, MN 55105.

Smorgasbord

The OECD published *Under Pressure: The Squeezed Middle Class* (April 2019; https://doi.org/10.1787/689afed1-en). "On average across OECD countries, the share of people in middle-income households, defined as households earning between 75% and 200% of the median national income, fell from 64% to 61% between the mid-1980s and mid-2010s. The economic influence of the middle class and its role as 'centre of economic gravity' has also weakened. The aggregate income of all middle-income households was four times the aggregate income of high-income households three decades ago; today, this ratio is less than three. . . . More than one-in-five middle-income households spend more than they earn. Over-indebtedness is higher for middle-income than for both low- and high-income households. . . . Middle-class lifestyle is typically associated with certain goods and

■ *Timothy Taylor is Managing Editor,* Journal of Economic Perspectives, *based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.33.3.249 doi=10.1257/jep.33.3.249

services and certain living conditions, such as decent housing, good education and good and accessible health services. However, the prices of core consumption goods and services such as health, education and housing have risen well above inflation, while middle incomes have been lagging behind. In particular, ageing and new medical technologies have driven up the cost of health services; the race for diplomas is pressing parents to invest more and more in education while, at the same time, education services became more costly in a number of countries; the geographical polarisation of jobs is pushing up housing prices in large urban areas, precisely where most rewarding jobs are available."

Steven A. Altman, Pankaj Ghemawat, and Phillip Bastian have written the *DHL Global Connectedness Index 2018: The State of Globalization in a Fragile World* (February 2019; https://www.logistics.dhl/content/dam/dhl/global/core/documents/pdf/glo-core-gci-2018-full-study.pdf). "Surprisingly, one commonality between globalization's supporters and its critics is that both tend to believe the world is *already* far more globalized than it really is. . . . The world is both *more globalized* than ever before and *less globalized* than most people perceive it to be. The intriguing possibility embodied in that conclusion is that companies and countries have far larger opportunities to benefit from global connectedness *and* more tools to manage its challenges than many decision-makers recognize." The report discusses a "survey of 6,035 managers across three advanced economies (Germany, the UK, and the US) and three emerging economies (Brazil, China, and India) that we conducted in 2017. On average, the managers guessed that the world was five times more deeply globalized than it really is! In fact, their perceptions were no more accurate than those of students surveyed across 138 countries or members of the general public in the United States. And CEOs and other senior executives had even *more* exaggerated perceptions than did junior and middle managers—perhaps because their own lives tend to be far more global than those of their employees and customers. . . . The combined output of all multinational firms outside of their home countries added up to only 9% of global economic output in 2017, and just 2% of all employees around the world worked in the international operations of multinational firms. . . . Most countries' international flows are so highly concentrated with key partner countries (usually neighbors) that it hardly makes sense to think of them as global at all. . . . Thus, despite the widespread perception that advances in transportation and telecommunications technologies are rendering distance irrelevant, international activity continues to be more intense among proximate countries."

The National Academies of Sciences, Engineering, and Medicine have published *A Roadmap to Reducing Child Poverty*, edited by Greg Duncan and Suzanne Le Menestrel (February 2019; https://www.nap.edu/catalog/25246/a-roadmap-to-reducing-child-poverty). "[M]any studies show significant associations between poverty and poor child outcomes, such as harmful childhood experiences, including maltreatment, material hardship, impaired physical health, low birthweight, structural changes in brain development, and mental health problems. Studies also show significant associations between child poverty and lower educational attainment, difficulty obtaining steady, well-paying employment in adulthood, and a greater likelihood of risky behaviors, delinquency, and criminal behavior in adolescence and adulthood. Because these correlations do not in themselves prove that low income is the active ingredient producing worse outcomes for children, the committee focused its attention on the literature addressing the *causal* impacts of childhood poverty on children. The committee concludes from this review that the weight of the causal evidence does indeed indicate that income poverty itself causes negative child outcomes, especially when poverty occurs in early childhood or persists throughout a large portion of childhood. . . . The committee also reviewed the much less extensive evidence on the macroeconomic costs of child poverty to measure how much child poverty costs the nation overall. Studies in this area attempt to attach a monetary value to the reduction in adult productivity, increased costs of crime, and increased health expenditures associated with children growing up in poor families. Estimates of these costs range from 4.0 percent to 5.4 percent of Gross Domestic Product—roughly between $800 billion and $1.1 trillion annually if measured in terms of the size of the U.S. economy in 2018. As we demonstrate below, outlays for new programs that would reduce child poverty by 50 percent would cost the United States much less than these estimated costs of child poverty."

Kevin L. Kliesen, Brian Levine, and Christopher J. Waller discuss "Gauging Market Responses to Monetary Policy Communication" (*Review*, Federal Reserve Bank of St. Louis, Second Quarter 2019, pp. 69–92; https://doi.org/10.20955/r.101.69-91). The authors point out that a century ago, an unofficial motto attributed to the Bank of England was "Never explain, never apologize." From 1967 to 1992, the main method of communication for the Federal Open Market Committee (FOMC) was to release a public statement 90 days after its meetings—not right after meetings. In contrast, "[t]he modern model of central bank communication suggests that central bankers prefer to err on the side of saying too much rather than too little. The reason is that most central bankers believe that clear and concise communication of monetary policy helps achieve their goals. . . . We find that Fed communication is associated with changes in prices of financial market instruments such as Treasury securities and equity prices. However, this effect varies by type of communication, by type of instrument, and by who is doing the speaking. . . . Perhaps not surprisingly, we find that the largest financial market reactions tend to be associated with communication by Fed Chairs rather than by other Fed governors and Reserve Bank presidents and with FOMC meeting statements rather than FOMC minutes."

Andreas Schrimpf and Vladyslav Sushko present "Beyond LIBOR: A Primer on the New Benchmark Rates" (*BIS Quarterly Review*, March 2019, pp. 29–52; https://www.bis.org/publ/qtrpdf/r_qt1903e.htm). "As of mid-2018, about $400 trillion worth of financial contracts referenced London interbank offered rates (LIBORs) in one of the major currencies. . . . A major impetus for reform comes from the need to strengthen market integrity following cases of misconduct involving banks' LIBOR submissions. To protect them against manipulation, the new (or reformed) benchmark rates would ideally be grounded in actual transactions and liquid markets rather than be derived from a poll of selected banks. . . . The reform process constitutes a major intervention for both industry and regulators, as it is akin to surgery on the pumping heart of the financial system. . . . The new risk-free rates (RFRs) provide for robust and credible overnight reference rates, well suited for many purposes and market needs. In the future, cash and derivatives markets are expected to migrate to the RFRs as the main set of benchmarks. . . . It is possible that, ultimately, a number of different benchmark formats will coexist, fulfilling a variety of purposes and market needs. The jury is still out on whether any resulting market segmentation would lead to material inefficiencies or could even be optimal under the new normal." This essay is a useful overview of what has happened since Darrell Duffie and Jeremy C. Stein wrote "Reforming LIBOR and Other Financial Market Benchmarks" in the Spring 2015 issue of this journal.

Collections of Essays

Meredith A. Crowley has edited *Trade War: The Clash of Economic Systems Threatening Global Prosperity*, a readable e-book of 11 essays (VoxEU.org, Centre for Economic and Policy Research Press, May 2019; available with free registration at https://voxeu.org/content/trade-war-clash-economic-systems-threatening-global-prosperity). From Crowley's introduction: "A trade war of unprecedented scope and magnitude currently engulfs the world's two largest economies—the US and China. . . . Multiple factors—the unprecedented economic growth of an economy operating outside the traditional Western capitalist model; new structures of production with supply chains spanning the globe; geographically concentrated job losses within the US; and a multilateral trading system that has stagnated and failed to keep pace with changes in the world economy—have all contributed to the current mess. The current problems extend well beyond the highly visible US–China conflict to the wider community of countries struggling with the interface between Chinese state capitalism and their own capitalist systems, the failure of the WTO to make progress with multilateral negotiations over almost anything, and a dispute resolution system that has veered off track. From our current vantage point, the prospects for the future of the multilateral trading system look grim. . . . Yet, in the middle of ongoing negotiations to resolve the US–China conflict, it is important to remember that the open, liberal multilateral trading system has delivered enormous benefits in its 75-year history—Ralph Ossa estimates the gains from trade amount to one-quarter of world income."

The American Statistical Association has devoted a special supplemental issue of its journal *The American Statistician* to the theme "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$" (vol. 73, no. S1, March 2019; https://www.tandfonline.com/toc/utas20/73/sup1). Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar contribute a useful overview essay, "Moving to a World beyond '$p < 0.05$.'" "We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term 'statistically significant' entirely. Nor should variants such as 'significantly different,' '$p < 0.05$,' and 'nonsignificant' survive, whether expressed in words, by asterisks in a table, or in some other way. Regardless of whether it was ever useful, a declaration of 'statistical significance' has today become meaningless. . . . In sum, 'statistically significant'—don't say it and don't use it." The special issue is then packed with 43 essays from a wide array of experts and fields that discuss what might follow if the language of statistical significance was eliminated.

Heather Boushey, Ryan Nunn, and Jay Shambaugh have edited a collection of eight essays on the subject *Recession Ready: Fiscal Policies to Stabilize the American Economy* (Hamilton Project, Brookings Institution and Washington Center for Equitable Growth, May 2019; http://www.hamiltonproject.org/papers/recession_ready_fiscal_policies_to_stabilize_the_american_economy). From their introduction: "[I]ncreasing the automatic nature of fiscal policy would be helpful. Increasing spending quickly could lead to a shallower and shorter recession. Using evidence-based automatic 'triggers' to alter the course of spending would be a more-effective way to deliver stimulus to the economy than waiting for policymakers to act. Such well-crafted automatic stabilizers are the best way to deliver fiscal stimulus in a timely, targeted, and temporary way. There will likely still be a need for discretionary policy; but by automating certain parts of the response, the United States can improve its macroeconomic outcomes." They mention the proposal from Claudia Sahm that when "the three-month moving average of the national unemployment rate has exceeded its minimum during the preceding 12 months by at least 0.5 percentage points," the federal government should have legislation in place that would immediately make a direct payment to adults of about of about 0.7% of GDP (which could be repeated later if the recession persists). Other chapters of the book consider specific programs that could be redesigned to increase automatically when a recession begins, including a transportation infrastructure plan, unemployment benefits, Temporary Assistance for Needy Families, the federal share of Medicaid and the Children's Health Insurance Program, and others. These essays supplement the three-paper "Symposium on Fiscal Policy" in the Spring 2019 issue of this journal.

The annual Conference on Research in Income and Wealth focuses on improved measurement of economic statistics. Katharine G. Abraham, Ron S. Jarmin, Brian Moyer, and Matthew D. Shapiro organized this year's conference, held March 15–16, 2019, in Bethesda, Maryland, on the theme of "Big Data for 21st Century Economic Statistics." Sixteen of the papers (and their presentation slides) are available at the website of the conference organizer, the National Bureau of Economic Research (https://papers.nber.org/sched/CRIWs19). A selection of some titles gives a flavor of the proceedings: "Re-engineering Key National Economic Indicators," by Gabriel Ehrlich, John C. Haltiwanger, Ron S. Jarmin, David Johnson, and Matthew D. Shapiro; "Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity," by Edward L. Glaeser, Hyunjin Kim, and Michael Luca; "Transforming Naturally Occurring Text Data into Economic Statistics: The Case of Online Job Vacancy Postings," by David Copple, Bradley J. Speigner, and Arthur Turrell; "From Transactions Data to Economic Statistics: Constructing Real-Time, High-Frequency, Geographic Measures of Consumer Spending," by Aditya Aladangady, Shifrah Aron-Dine, Wendy Dunn, Laura Feiveson, Paul Lengermann, and Claudia R. Sahm; and "Valuing Housing Services in the Era of Big Data: A User Cost Approach Leveraging Zillow Microdata," by Marina Gindelsky, Jeremy Moulton, and Scott A. Wentland. This research complements the three-paper "Symposium on Public Provision of Economic Data" in the Winter 2019 issue of this journal.

The Harvard Project on Climate Agreements and Harvard's Solar Geoengineering Research Program have collaborated to publish *Governance of the Deployment of Solar Geoengineering*, an introduction followed by 26 short essays (November 2018; https://www.c2g2.net/wp-content/uploads/Harvard-Project-Solar-Geo-Governance-Briefs-181126.pdf). In "The Implications of Uncertainty and Ignorance for Solar Geoengineering," Richard J. Zeckhauser and Gernot Wagner write, "Risk, uncertainty, and ignorance are often greeted with the precautionary principle: 'do not proceed.' Such inertia helps politicians and bureaucrats avoid blame. However, the future of the planet is too important a consequence to leave to knee-jerk caution and strategic blame avoidance. Rational decision requires the equal weighting of errors of commission and omission. . . . That also implies that the dangers of SG [solar geoengineering]—and they are real—should be weighed objectively and dispassionately on an equal basis against the dangers of an unmitigated climate path for planet Earth. The precautionary principle, however tempting to invoke, makes little sense in this context. It would be akin to suffering chronic kidney disease, and being on the path to renal failure, yet refusing a new treatment that has had short-run success, because it could have long-term serious side effects that tests to date have been unable to discover. Failure to assiduously research geoengineering and, positing no red-light findings, to experiment with it would be to allow rising temperatures to go unchecked, despite great uncertainties about their destinations and dangers. That is hardly a path of caution."

Economists Speak

Christopher J. Ruhm delivered the "Presidential Address: Shackling the Identification Police?" to the Southern Economic Association (*Southern Economic Journal*, vol. 85, no. 4, April 2019, pp. 1016–26; https://onlinelibrary.wiley.com/doi/abs/10.1002/soej.12333). "To summarize, clean identification strategies will frequently be extremely useful for examining the partial equilibrium effects of specific policies or outcomes—such as the effects of reducing class sizes from 30 to 20 students or the consequences of extreme deprivation in-utero—but will often be less successful at examining the big 'what if' questions related to root causes or effects of major changes in institutions or policies. . . . Have the identification police become too powerful? The answer to this question is subjective and open to debate. However, I believe that it is becoming increasingly difficult to publish research on significant questions that lack sufficiently clean identification and, conversely, that research using quasi-experimental and (particularly) experimental strategies yielding high confidence but on questions of limited importance are more often being published. In talking with PhD students, I hear about training that emphasizes the search for discontinuities and policy variations, rather than on seeking to answer questions of fundamental importance. At professional presentations, experienced economists sometimes mention 'correlational' or 'reduced-form' approaches with disdain, suggesting that such research has nothing to add to the canon of applied economics."

David A. Price interviews R. Preston McAfee (*Econ Focus*, Federal Reserve Bank of Richmond, Fourth Quarter 2018, pp. 18–23; https://www.richmondfed.org/publications/research/econ_focus/2018/q4/interview). "First, let's be clear about what Facebook and Google monopolize: digital advertising. The accurate phrase is 'exercise market power,' rather than monopolize, but life is short. Both companies give away their consumer product; the product they sell is advertising. While digital advertising is probably a market for antitrust purposes, it is not in the top 10 social issues we face and possibly not in the top thousand. Indeed, insofar as advertising is bad for consumers, monopolization, by increasing the price of advertising, does a social good. . . . That leaves . . . two places where I think we have a serious tech antitrust problem. . . . My concern is that phones, on which we are incredibly dependent, are dominated by two firms that don't compete very strongly. While Android is clearly much more open than Apple, and has competing handset suppliers, consumers face switching costs that render them effectively monopolized. . . . The second place I'm worried about significant monopolization is Internet service. In many places, broadband service is effectively monopolized. . . . I'm worried about that because I think broadband is a utility. You can't be an informed voter, you can't shop online, and you probably can't get through high school without decent Internet service today. So that's become a utility in the same way that electricity was in the 1950s. Our response to electricity was we either did municipal electricity or we did regulation of private provision. Either one of those works. That's what we need to do for broadband."

Discussion Starters

Peter Cappelli thinks "Your Approach to Hiring Is All Wrong" (*Harvard Business Review*, May–June 2019; https://elb.hbr.org/2019/05/recruiting#your-approach-to-hiring-is-all-wrong). "Businesses have never done as much hiring as they do today. They've never spent as much money doing it. And they've never done a worse job of it. . . . The recruiting and hiring function has been eviscerated. Many U.S. companies—about 40%, according to research by Korn Ferry—have outsourced much if not all of the hiring process to 'recruitment process outsourcers,' which in turn often use subcontractors, typically in India and the Philippines. . . . Survey after survey finds employers complaining about how difficult hiring is. . . . But clearly they are hiring much more than at any other time in modern history, for two reasons. The first is that openings are now filled more often by hiring from the outside than by promoting from within. In the era of lifetime employment, from the end of World War II through the 1970s, corporations filled roughly 90% of their vacancies through promotions and lateral assignments. Today the figure is a third or less. When they hire from outside, organizations don't have to pay to train and develop their employees. . . . The second reason hiring is so difficult is that retention has become tough: Companies hire from their competitors and vice versa, so they have to keep replacing people who leave. Census and Bureau of Labor Statistics data shows that 95% of hiring is done to fill existing positions. Most of those vacancies are caused by voluntary turnover. . . . The root cause of most hiring, therefore, is drastically poor retention."

Jeremiah Dittmar and Skipper Seabold explain how "Gutenberg's Moving Type Propelled Europe towards the Scientific Revolution" (*LSE Business Review*, March 19, 2019; https://blogs.lse.ac.uk/businessreview/2019/03/19/gutenbergs-moving-type-propelled-europe-towards-the-scientific-revolution/). "Printing was not only a new technology: it also introduced new forms of competition into European society. Most directly, printing was one of the first industries in which production was organised by for-profit capitalist firms. These firms incurred large fixed costs and competed in highly concentrated local markets. Equally fundamentally—and reflecting this industrial organisation—printing transformed competition in the 'market for ideas'. Famously, printing was at the heart of the Protestant Reformation, which breached the religious monopoly of the Catholic Church. But printing's influence on competition among ideas and producers of ideas also propelled Europe towards the scientific revolution. . . . Following the introduction of printing, book prices fell steadily. The raw price of books fell by 2.4 per cent a year for over a hundred years after Gutenberg. Taking account of differences in content and the physical characteristics of books, such as formatting, illustrations and the use of multiple ink colours, prices fell by 1.7 per cent a year. . . . Printing provided a new channel for the diffusion of knowledge about business practices. The first mathematics texts printed in Europe were 'commercial arithmetics', which provided instruction for merchants. With printing, a business education literature emerged that lowered the costs of knowledge for merchants. The key innovations involved applied mathematics, accounting techniques and cashless payments systems." For a detailed discussion, see the authors' research paper, "New Media and Competition: Printing and Europe's Transformation after Gutenberg" (Centre for Economic Performance Discussion Paper 1600, January 2019; http://cep.lse.ac.uk/pubs/download/dp1600.pdf).

Michael Manville offers a discussable counterfactual in "Longer View: The Fairness of Congestion Pricing: The Choice Between Congestion Pricing Fairness and Efficiency Is a False One" (*Transfers*, Spring 2019, pp. 1–6; https://transfersmagazine.org/longer-view-the-fairness-of-congestion-pricing/). "Suppose we had a world where all freeways were priced, and where we used the revenue to ease pricing's burden on the poor. Now suppose someone wanted to change this state of affairs, and make all roads free. Would we consider this proposal fair? The poorest people, who don't drive, would gain nothing. The poor who drive would save some money, but affluent drivers would save more. Congestion would increase, and so would pollution. The pollution would disproportionately burden low-income people. With priced roads, poor drivers were protected by payments from the toll revenue. With pricing gone, the revenue would disappear as well, and so would compensation for people who suffered congestion's costs. This proposal, in short, would reduce both efficiency *and* equity. It would harm the vulnerable, reward the affluent, damage the environment, and make a functioning public service faulty and unreliable. . . . We have so normalized the current condition of our transportation system that we unthinkingly consider it fair and functional. It is neither. Our system is an embarrassment to efficiency and an affront to equity."

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary access to JEP articles, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $22.00, $32.00, or $42.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

The Journal of

Economic Perspectives

Summer 2019, Volume 33, Number 3

Symposia

Markups

Issues in Antitrust

Articles

Features

AMERICAN
ECONOMIC
ASSOCIATION