The Journal of Economic Perspectives

A journal of the American Economic Association

Summer 2022

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 2022, Vol. 36, 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 $24.00, $34.00, or $44.00 depending on income; for an additional $15.00, you can receive this journal in print. The journal is freely available online. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 36 • Number 3 • Summer 2022

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 36, Number 3—Summer 2022—Pages 3–28

Intangible Capital and Modern Economies

Carol Corrado, Jonathan Haskel,
Cecilia Jona-Lasinio, and Massimiliano Iommi

Every practitioner of economics, whether student or professional, will at some point be asked about growth, innovation, and economic performance. Why are many African countries so poor? How did Japanese car companies come to dominate world production? How did some firms like Walmart, Amazon, and Facebook get to be so big, while others like Sears, Webvan, and MySpace crashed and burned? At this point, the practitioner will recall the textbook account of a production function, in which output is a function of inputs of capital, labor, and technology. To make the abstractions concrete, most textbooks are peppered with examples from agriculture or manufacturing. In agriculture, capital is tractors, labor is farmworkers, and technology is the "know-how" of crop production. Technology or know-how is discussed broadly as advancing through research and development (R&D) and policies that govern the protection of patents and other forms of intellectual property.

■ *Carol Corrado is Distinguished Principal Research Fellow in Economics, The Conference Board, New York City, New York, and Senior Policy Scholar, Center on Business and Public Policy, Georgetown University McDonough School of Business, Washington, DC. Jonathan Haskel is Professor of Economics, Imperial College Business School, and an External Member of the Monetary Policy Committee, Bank of England, London, United Kingdom. Cecilia Jona-Lasinio is Senior Researcher at the Italian Statistical Institute (ISTAT) and Professor of International Economics at LUISS University, both in Rome, Italy. Massimiliano Iommi is a Senior Researcher, Italian Statistical Institute, and a Research Fellow, LUISS Lab of European Economics, both in Rome, Italy. Their email addresses are carol.corrado@conferenceboard.org, j.haskel@imperial.ac.uk, Cecilia Jona-Lasinio cjonalasinio@luiss.it, and Massimiliano Iommi miommi@luiss.it.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.3.

Table 1
The World's Largest Companies by Market Capitalization, March 31, 2021
(billions of US dollars)

Company name	*Market capitalization*	*Tangible assets*	*R&D assets*
Apple	2,051	344	75
Saudi Aramco	1,920	322	5
Microsoft	1,778	245	92
Amazon	1,558	330	137
Alphabet	1,393	300	105
Facebook	839	141	51

Source: PWC and company reports (market capitalization and tangible assets for 2020). R&D assets are authors' estimates of 2020 R&D stock based on time series of R&D spending from company reports.

But when this practitioner of economics encounters the real world, this basic production function approach exhibits some glaring holes. Table 1 sets out the world's leading companies by market capitalization in March 2021. Market capitalization refers to the total value of the company, based on stock market valuations. (It should be noted that some companies, like Saudi Aramco, remain primarily owned by the government of Saudi Arabia.) A first lesson from Table 1 is that the value of these companies is clearly not based on the textbook physical or "tangible" capital, which covers "property, plant, and equipment." The gap between tangible assets as reported in corporate annual reports and the market value of these companies is enormous, even though tangible assets do include, for example, Amazon's property, plant, and equipment in cloud server farms.

Perhaps then the value of these companies is more closely related to their "intangible" assets, that is, their "know-how"? The final column of Table 1 sets out an estimate of the capitalized value of spending on research and development by these companies, based on calculations by the authors that sum the value of past R&D spending by the firms and assume a depreciation rate of 15 percent. However, combining these figures with tangible assets does little to explain market capitalization.

In what follows, we will argue that understanding modern firms and indeed modern economies requires broadening the concept of capital beyond tangible assets to include intangibles, and that research and development spending is not the only way to capture intangible capital. Indeed, R&D spending is extraordinarily skewed by size of firm and by industry. The OECD (2017) reports, "In 2014, the top 10 percent of [the world's largest] corporate R&D investors (i.e., the top 200 companies with their affiliates) accounted for about 70 percent of R&D expenditure and 60 percent of . . . inventions patented in the [world's] five top IP offices." In the US economy, just four industry groups—chemicals, computer and electronic products, transportation equipment, and information services—accounted for

more than 70 percent of R&D performed in 2018.[1] Many substantial industries including retail, finance, and most professional services—say, companies such as LinkedIn—do little or no R&D or patenting. But innovative firms do invest in other types of knowledge not classified as R&D: software tools, attributed designs, and strategies for improving brand awareness, business practices, services delivery, or managing after-sale services, and others.[2]

In what follows, we shall discuss intangible capital as reflecting investments in many types of knowledge-based, nonphysical assets. We begin by discussing what constitutes investment in knowledge-based assets and how accounting for such assets reshapes our thinking about macroeconomic data on investment. We then turn to issues of how intangible capital relates to growth theory and practical growth accounting. We consider how the growth and ownership of intangible capital may affect competitiveness across firms. We lay out some of the challenges underlying measurement of intangible capital and discuss how it affects estimates of productivity in the US and European economies in recent decades. Finally, we address the conundrum of why, despite a growth in intangible capital and what seems to be a modern technological revolution, productivity growth has slowed down since the Great Recession.

What Is Intangible Investment?

The potential importance of intangible investment in understanding the economy has deep roots in economics. For example, in the 1970s and 1980s, there were efforts to treat research and development as an intangible capital asset in both firm-level growth and neoclassical growth studies (Griliches 1973, 1979, 1986). The academic thinking about brand as strategic capital of the firm is rooted in the management/marketing literature that developed somewhat later (Farquhar 1989; Aaker 1991). But the significance of intangible investments in the structure of organizations and the macroeconomy did not emerge until the information technology-driven productivity "boom" of the late 1990s (Brynjolfsson and Hitt 2000; Brynjolfsson, Hitt, and Yang 2002). That boom was accompanied by a large widening gap between market valuation of firms based on equity markets and accounting valuations of firms based on the physical plant, property, and equipment—that is, gaps such as those shown in Table 1. Influential research from accounting underscored

[1] Based on our own calculations using figures reported by the US National Science Foundation at https://ncses.nsf.gov/pubs/nsf20316/.

[2] International standards for R&D surveys are set out in the Frascati Manual 2015 (OECD 2015) subtitled, *The Measurement of Scientific, Technological and Innovation Activities*. It defines R&D as activity that comprises "creative and systematic work undertaken to increase the stock of knowledge . . . [and] to devise new applications of available knowledge." R&D expenditure survey respondents are typically instructed to not include expenditures on efficiency surveys; management or organization studies; marketing research and consumer surveys; advertising or promotions; the payment for another's patent, model, production, or process; prospecting or exploration for natural resources; or research in connection with literary, historical, or similar projects (Moris 2018).

that brand names, new products, and intangible assets such as software-enabled procurement systems were key drivers of the financial outcomes of many of the nation's most innovative companies (Lev 2001). Indeed, Lev (2005) suggested that company reports consider new products/services development, customer relations, human resources, and organizational capital as assets. These observations and findings spurred measurement-oriented economists to pursue the notion that there was more to business investment than captured in standard macroeconomic measures (for example, Young 1998; Nakamura 1999, 2001).

The approach of Corrado, Hulten, and Sichel (2005, 2009) as summarized in Table 2 built upon these works. Their intangible assets approach expands the range of spending by firms that should be viewed as an investment. It applies a fundamental economic criterion that defines investment, namely, that business (or public) investments are outlays expected to yield a return in a future period.

The principle obviously applies to tangible spending and to research and development spending: for example, spending on a tractor or a robot is an investment, and so is R&D that yields a drug formula and software code that (say) guides a delivery truck more efficiently. In an economic sense, investments in industrial design, market development, employee training, organizational change, and even songs and film scripts likewise provide ongoing revenue. The categorization of intangible investment proposed by Corrado, Hulten, and Sichel (2005, 2009) suggests a wide class of intangible assets, from databases to business processes. The intangible assets listed in Table 2 are attractive for understanding the market capitalization of the companies in Table 1 because those companies tend to be based on software, data, design, operations networks, and brand.

The OECD (2013) has adopted the taxonomy in Table 2, using "knowledge-based capital" to describe it. The European Union, which since 2003 commissioned a series of studies of productivity accounts known as EU KLEMS—where the acronym stands for inputs of capital (K), labor (L), energy (E), materials (M), and services (S)—includes in its most recent version the complete list of intangible assets from Table 2 via an INTANProd production module for each country.[3]

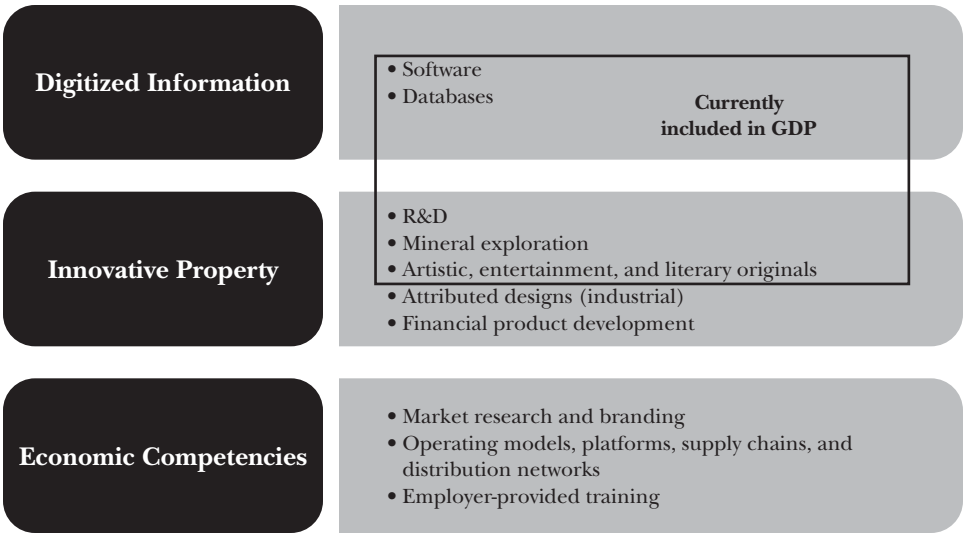
Intangibles in Existing Data

To what extent do official macroeconomic and financial data incorporate intangible capital? The incorporation of intangibles into national accounts is moving, but slowly; their incorporation into company financial accounts has not progressed materially, and as matters now stand, the treatment of intangibles is conceptually inconsistent (for a recent self-assessment, see CPA Ontario 2021).

In official calculations of GDP, there has been a relatively recent recognition of certain intangible assets including R&D, mineral exploration, computer software (blended with internally produced databases), and entertainment, artistic, and literary originals—the assets "boxed" in Table 2. GDP arbiters have been hesitant

[3]The EU KLEMS & INTANProd database is available from the LUISS Lab of European Economics at LUISS University (https://euklems-intanprod-llee.luiss.it/).

Table 2
Intangible Capital: Broad Categories and Types of Investment



Source: Authors' elaboration of Corrado, Hulten, and Sichel (2005, 2009).

to embrace the idea that the asset boundary of an organization encompasses intangible investments in industrial design, marketing and branding, management practices, and employer-provided training—the complete Table 2 approach—for some reasons we elaborate on below.

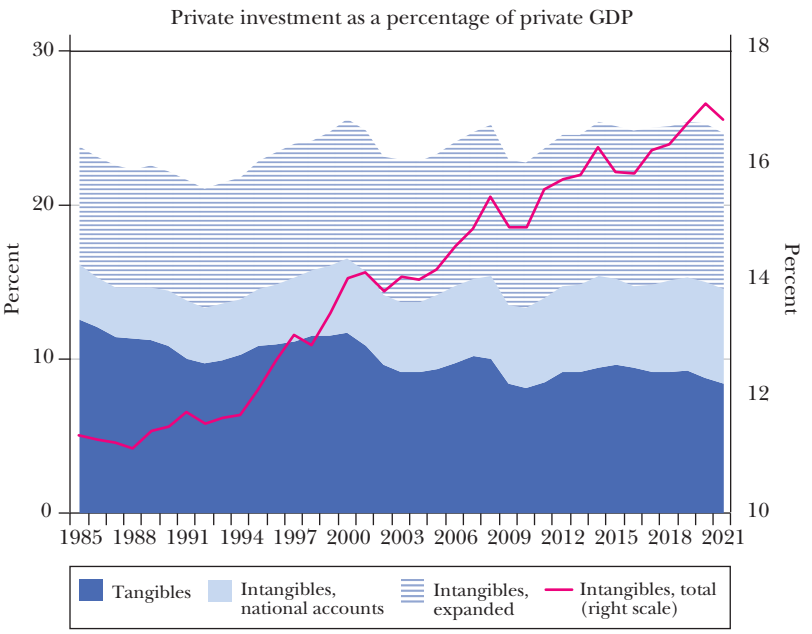
The standards for reporting intangibles into company accounts are problematic and asymmetric. For example, International Accounting Standard #38 on "Intangible Assets"[4] generally disallows the capitalization of most internally generated intangible assets, like most R&D, software, and brand/organization development costs, but it allows capitalization of externally generated intangible assets like patent portfolios and customer lists when acquired via merger activity. Researchers who use values for intangible assets on firms' balance sheets should be aware that they largely arise from acquisitions, not from production, thus creating a situation in which changes in reported assets do not reflect actual investment flows in an economy.

Intangible Investment in the Macroeconomy

Implementation of an expanded framework for investment and intangible capital provides a new view on the characteristics and performance of the

[4]Available at https://www.ifrs.org/issued-standards/list-of-standards/ias-38-intangible-assets/.

Figure 1
Rates of Private Nonresidential Investment in the United States, Tangible and Intangible, 1987 to 2021



Source: Authors' elaboration of data on investment by broad category from the US national accounts and US intangibles module of EU KLEMS & INTANProd.
Note: GDP includes all intangible investment.

macroeconomy. Figure 1 shows rates of private nonresidential intangible and tangible investment based on this framework for the US economy since 1985. Following Table 2, it separates intangible investment that is included in national accounts from the whole. The rate for tangible investment, the dark shaded portion at the bottom, drifts down 4 percentage points over the period shown, from about 12½ percent of private sector GDP in 1985 to about 8½ percent in 2021. Total investment in the economy, which adds investment in intangibles and is shown by the sum of the shaded areas, edges up by more than 1 percentage point, driven by growth in the relative importance of intangible investment. Indeed, the rate of total intangible investment (plotted separately as a line on the right scale) rises rather dramatically over the period shown and now stands at about 16¾ percent of GDP.

Another message from Figure 1 is that total investment in intangibles in the United States substantially exceeds components included in official statistics; the same can be said for the major economies of Europe. Practitioners analyzing macroeconomic trends, who may have been taught that research and development is a sufficient proxy for innovation effort, should be aware of the relative magnitudes displayed in Figure 1. Regarding private R&D, in cross-country data covering

selected countries in Europe and the United States (described below), the correlation between growth in R&D capital and total intangible capital excluding R&D is 0.32. The correlation between the official components of intangible capital and the expanded components is 0.28. These correlations suggest that much is missing in official macroeconomic data on private investment.

Although the primary focus of this paper is on how intangible capital affects growth and competitive mechanisms in economies, some preliminary work suggests that the rise of intangible capital as a strategic factor input also has the potential for altering cyclical patterns. This includes patterns of investment and factor input demands, and perhaps the responsiveness of inflation to economic conditions in the short run. Research on the formulation of investment demand argues that intangibles are less sensitive to changes in interest rates than tangibles, reflecting their higher user cost and tendency to be less reliant on secured debt financing (for example, Crouzet and Eberly 2019; Haskel and Westlake 2018, chapter 8; Döttling and Ratnovski 2020). Figure 2 displays fluctuations in the intellectual property products (that is, intangibles already included in national accounts) as a share of private nonresidential investment, using quarterly data for the United States. Notice that during the recession periods (shown as shaded bars) these investments tend to keep rising, which suggests that these investments are the last category of capital spending cut during downturns.

Businesses may view the acquisition of software (and other intangibles) as moves to increase efficiency that dampen the impact of workforce layoffs and cutbacks in customer demand. The fact that intangible capital increasingly reflects knowledge built from the analysis of data likely explains the recent persistence of its relative strength; as an example, half of the respondents in global survey of companies administered by McKinsey & Company reported that the pandemic-induced economic downturn had no effect on their investments in artificial intelligence, while 27 percent reported increasing them (Zhang et al. 2021, p. 103). An implication is that intangible capital (or some forms of it) may help firms to adjust production relatively rapidly to changes in economic conditions.

Intangible Capital and Growth Theory

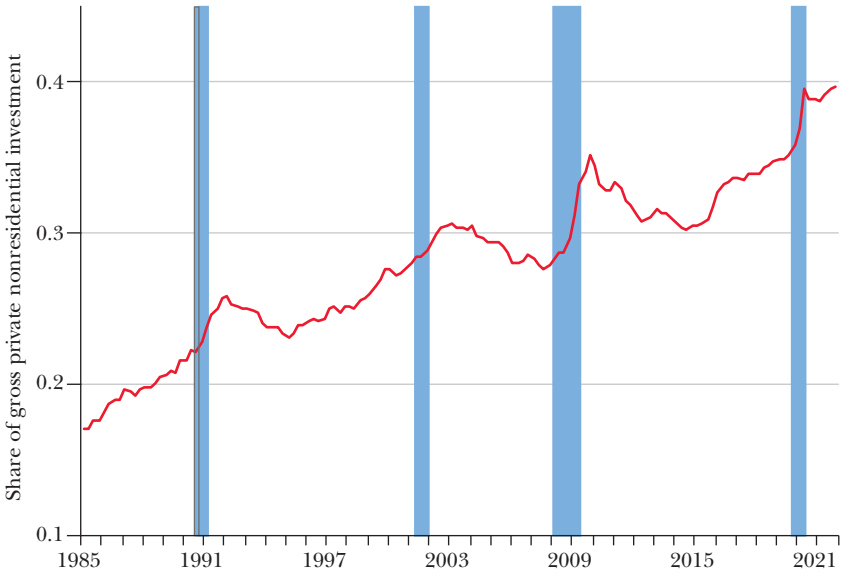
Here, we review how the intangibles approach relates to neoclassical and endogenous growth theory, as well as to strands of literatures concerning human capital, management, and business innovation.

How Does the Intangible Capital Approach Fit into Growth Theory?

The standard production function begins with output as a function of (quality-adjusted) inputs of capital and labor. The empirics of standard growth accounting are a powerful macroeconomic tool, so it is useful that the intangible capital framework augments standard growth accounting, rather than seeking to replace it.

One way of understanding how the intangible investment framework builds on the standard aggregate production function is to recognize that, as Milgrom

Figure 2
Intellectual Property Products Investment in the United States, 1985:I–2021:IV



Source: Authors' elaboration of quarterly data from the national income and product accounts.
Notes: Intellectual property products include software, R&D, and entertainment originals. Shaded areas are periods of business recession as defined by the National Bureau of Economic Research.

and Roberts (1990) have argued, the standard approach says nothing about what might be called the "coordination activities" within a firm required in production. A standard macroeconomic production function approach describes a set of possible production plans, but is not explicit about the costs of coordinating or managing their combination. Evaluating alternative plans, managing supply chains, and balancing competing interests in an organization is costly. If such costs are integral to generating ongoing returns, then such costs are investments. The intangibles approach accounts for "coordination activities" as long-lived investments in process efficiencies by grouping spending on new or reorganized business models under the heading of investment in organizational capital.

The standard production function approach also says little about "informative activities" that build long-lived demand, activities like marketing, market research, customer development, product promotion, and brand-building. The recognition of these activities as investment captures the insight that customer adoption of new products and new technologies typically is far from costless. Instead, such investments expand demand (and thus productive capacity) rather than change the production process (Hulten 2010, 2011). Though the introduction of demand-side considerations in growth analysis is a substantial departure from the neoclassical

and endogenous growth paradigms, it leads to considering how accounting for intangible capital affects the analysis of market power and imperfect competition, a point to which we will return below.

Measuring and accounting for this broader notion of intangible capital in fact provides a bridge to Romer's (1990) endogenous growth theory. In his approach, the aggregate production function has (implicitly) constant returns to "objects" like capital and labor but adds "ideas" and the potential for increasing returns to ideas. Jones (2019, pp. 864–5) elaborates:

> Whereas Solow divided the world into capital and labor, Romer makes a more basic distinction: between ideas, on the one hand, and everything else (call them "objects") on the other. Objects are the traditional goods that appear in economics, including capital, labor, human capital, land, highways, lawyers, a barrel of oil, a bushel of soybeans . . . An idea is a design, a blueprint, or a set of instructions for starting with existing objects, and transforming or using them in some way. . . . Examples include calculus, the recipe for a new antibiotic, Beethoven's Fifth Symphony, the design of the latest quantum computer.

The source of the increasing returns to ideas is their nonrival property. Romer (1990) illustrated this property with the example of oral rehydration therapy. This simple formula, essentially requiring a packet of sugar, salt, and potassium to be mixed with water, cures diarrhea and has saved literally millions of lives in developing countries. Suppose there is one plant in the world producing such packets. If a rival set up an identical plant, what inputs would be needed to produce the same number of packets? Romer's insight was that a new firm would need to employ a second set of machines and workers but could freely use the existing "idea"—the formula for the treatment—because it's available on Wikipedia and would not have to be invented anew.[5] In this sense, the production function has constant returns to objects but increasing returns to ideas.

Returning to Table 1, consider trying to duplicate the "ideas" that are Apple. Until 2008, the leading cell phone manufacturer was the Finnish company Nokia. Their phones were among the first to have auto-correct texting, Wi-Fi connections, and games. Yet with the introduction of Apple's iPhone, Nokia's market share collapsed. The Apple smartphone featured innovation like a touch screen technology and an aesthetic design—ideas that could, at least in principle, be licensed or copied by rivals. But Apple also had remarkably efficient supply chain management. When Nokia launched new products, customers waited for months to acquire them, whereas Apple could provide millions of new phones essentially on launch day (Cuthbertson, Furseth, and Ezell 2015). Apple's supply chain management knowledge cannot be copied from Wikipedia, and Apple's brand and reputation for service and delivery, while in public view, cannot be "shared" or copied in the

[5] See https://en.wikipedia.org/wiki/Oral_rehydration_therapy.

same way as a recipe for oral rehydration therapy. A firm seeking to be the same as Apple would have to invest in capital and labor, but also to invest in the knowledge that constitutes Apple's supply chain and its reputation for product/service quality. Nokia could not "freely utilize" this kind of knowledge to duplicate Apple.

The nonrival nature of intangibles is of course important, but the fact that intangible assets are partially appropriable takes center stage in the intangible capital approach. Without some degree of appropriability, there are no incentives for private business to invest in innovation, and without potential for commercial use, to paraphrase a comment attributed to Thomas Edison, there is no value in an idea. Consistent with endogenous growth theory, however, economies with investments in intangibles should still display increasing returns to those investments. But the intangible capital approach holds that the phenomenon does not just apply to investments in R&D; the potential for knowledge spillovers also extends to investments in business models, marketing strategies, and industrial design (among other areas) in models of intangible capital.

Whether the knowledge spillovers and knowledge stocks related to intangible capital should be termed "ideas" or not is mostly a semantic argument. In any case, appropriable knowledge stocks are termed "intangible capital" in the approach using a production function written as $Y = A\,F(L, K, R)$, where A is the technology that applies to the entire production function F. Here, intangible capital R is an input to the production process with several relevant traits: (a) it provides a flow of enduring income-generating services (and so is capital and not an intermediate); (b) more of it may be required along with more capital K and labor L to avoid diminishing marginal returns; but (c) as R is fundamentally nonrival, there is potential for increasing returns as the innovations embodied in intangible investments diffuse across firms, industries, and economies.

Why Isn't Intangible Capital Just Part of Labor?

One concern over expanding the conventional notion of business investment is to argue that much of what we have described will be captured by human capital, and in particular the talents of managers, engineers, and designers, which are accounted for in labor input. Does adding intangible capital pose a risk of double-counting?

The issue boils down to ownership of (or command of) the insights and intellectual property the managers and others are paid to develop. When Apple's founder and chief executive officer Steve Jobs passed away in 2011, the value of Apple did not disappear. Rather, a large part of his value was embodied in Apple itself. Formal studies of executives who leave famous companies, such as GE, find that they are often unable to repeat that success in other corporations, suggesting that they do not carry the corporate knowledge they created with them (Groysberg, McLean, and Nohria 2006).[6] Additionally, studies based on linked employer-employee data suggest that the marginal revenue product of managers exceeds their compensation,

[6]Formal studies of the value of a firm when the owner dies tend to find small effects in large firms, but larger effects for smaller firms; see the discussion in Smith et al. (2019, p. 1722).

and Piekkola (2016) even finds magnitudes in line with Figure 1's estimates of organizational capital generated within firms.

The human capital created by employer-provided training is a related concern, but studies demonstrate that firm-specific training (like the apprenticeships discussed in Zwick 2007) generates net returns to the firm, over and above the costs of the training and additional wages paid to employees with enhanced skills. Thus, it seems plausible that the skills embodied in the business practices of a firm are in several ways separable from the individuals working at the firm.

Intangible Capital, Competition, and GDP

If a firm is to use and pay for intangible capital, the capital must be produced and its owners rewarded. How do the production of intangible assets and the accompanying flows of reward fit into an overall vision of the economy?

A Model of an Economy

In a simplified model of the economy, production activity can be divided into two parts: 1) an "upstream" or "innovation" sector that produces ideas that can be commercialized, like a new system for organizing production or a software program adapted to the needs of the organization; and 2) a "downstream" or "production" sector that uses the knowledge generated by the upstream sector to produce final output.[7] By "final" output we mean output for sale to consumers or for export or investment: for simplicity, we ignore intermediate inputs. Figure 3 depicts these two interlinked production functions.

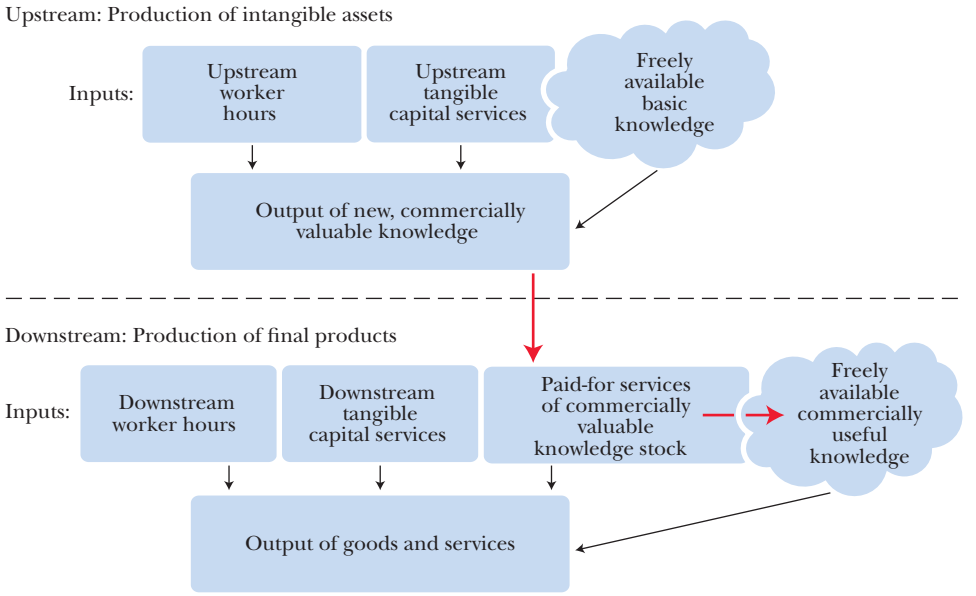
The outstanding stock of intangible capital in this framework, which might also be called "commercial knowledge," reflects the accumulation of upstream output, after adjusting for losses due to aging (the equivalent of depreciation). The production sector acquires commercial knowledge much as it acquires plant and equipment, via capital expenditure. But the stock of this knowledge is non-rival and only partially appropriable. The possible leakage from paid-for commercial knowledge to freely available useful knowledge is shown by the dotted arrow in the downstream sector.

The idea that innovators hold only temporary product market power for their inventions is a common feature of economic models of innovation. Such market power lasts for the time during which the innovator can sell or rent the knowledge for a monopoly price to the downstream sector, who in this framework is treated as a price-taker for knowledge.[8] We assume that prices for other inputs are competitive;

[7] The approach discussed here is based on Corrado, Hulten, Sichel (2005, 2009) as adapted in Corrado, Goodridge, and Haskel (2011).

[8] In this model the asset price for purchasing permanent use of commercial knowledge and the price of using this knowledge for a pre-set period of time (like a year) are linked via the Jorgenson (1963) user cost expression.

Figure 3
Conceptual Framework



Source: Authors' illustration.

final product prices are also competitive (given the cost of producing new commercial knowledge).

In contrast to this commercialized knowledge, "basic" knowledge, generated (say) via public funds for basic scientific research to universities, is assumed to be a free input in the upstream production function. Thus, while basic knowledge is an input to the production of commercial knowledge, it receives no factor payments to because its services are assumed to be freely available. "Basic" knowledge in this model is not viewed as stemming solely from scientific breakthroughs, though investments in branding and marketing, organization structure, and employer-provided training have long been modeled as complements with information technology equipment, as in Brynjolfsson and Hitt (2000); Corrado, Haskel, and Jona-Lasinio (2016) find justification for this approach in cross-country macroeconomic data.

This model's depiction of the two sectors captures some important aspects of business innovation in modern economies. The upstream sector would include firms that are almost fully reliant on the production of innovations in the form of new intangible assets—say, biotech startups producing new formulas for drugs—with the downstream sector comprising producers that acquire the use of the innovations via outright purchase or license agreements with annual payments. More generally, many innovating firms have their own internal "innovation labs" and "business

strategy teams" that produce and commercialize new ideas for downstream production (for example, Alphabet's "X" research arm). In our model, these innovation labs and strategy teams are then upstream knowledge producers residing within larger organizations, and the internal payments to these innovation labs and strategy teams represent their contribution to total revenue. This depiction of innovation is not limited to production of new technologies. For example, consider the downstream firm Peloton, which wishes to purchase the rights to music that can be played while people exercise. The firm can make "rental" payments to musicians for use of music in Peloton video exercises (now around 3 cents per song, as reported by Pahwa 2021), or the company could pay for the right to use a song (legally) forever.

Further, the intuition of an upstream entity commercializing knowledge helps, we believe, relate economic theory and measurement to the interests of management and innovation scholars. Such scholars typically find the economist's use of total factor productivity to represent innovation hard to reconcile with their detailed and diverse case studies of the internal process by which firms develop new products and processes whereas the innovation divisions with firms ("skunk works") described by Greenstein (2016) are, collectively, upstream sector knowledge producers in our model.

Dynamic innovative economies will maintain a continuous flow of differentiated innovations via investments in intangible capital. In the long run, firms will compare the after-tax returns to investments in innovation that build intangible capital with the returns to alternative long-term investments that build tangible capital. In this setting, non-zero innovator profits can persist, manifest as higher prices for intangible assets; for further discussion, see Corrado et al. (2022).

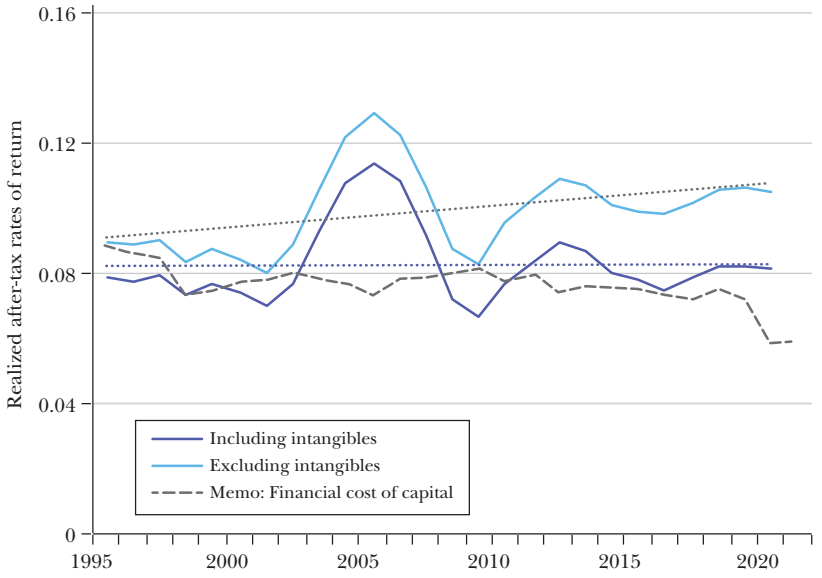
Implications for Measuring Market Power

The rewards that accrue to intangible investments are a part of business profits whether or not intangibles are measured or included in GDP, income, and fixed asset accounting. But if investments in intangible capital are not included, economies may appear to have abnormally high profits relative to the (mismeasured) capital employed—in fact, the higher the (uncounted) intangible investment, the greater the misperception. Using the investment data on tangible and intangible investment underlying the earlier Figure 1, Figure 4 shows that the after-tax rate of return implied by macroeconomic data is dramatically affected when investment is expanded to cover intangibles.

Figure 4 also shows a market-based cost of capital, calculated as a weighted average of the expected return on stocks and after-tax cost of debt. The gap between the financial cost of capital and realized (actual) returns to capital can serve as an indicator of market power, akin to the price markup discussed in much recent competition literature.

The message suggested by the (erroneous) gap based on the rate of return excluding intangibles is akin to the price markup calculated using (most) microdata sources. Firm-level databases based on company financial reports and/or microdata from official production surveys do not account for intangibles—and

Figure 4
Realized After-Tax Rates of Return, US Private Industries, 1995 to 2020



Source: Authors' elaboration of data from US national accounts, BEA, BLS, EU KLEMS & INTANProd, Federal Reserve financial accounts, and FRED databases.
Notes: Private industries exclude residential real estate and private households. Private capital includes fixed assets, inventories, and nonresidential land. Intangibles covers all assets listed in Table 2. The last point plotted for the financial cost of capital is 2021.

even miss the software and R&D components included in macrodata—and thus are difficult sources for depicting competitive developments accurately.

Implications for GDP and Growth Accounting

In the model of the economy depicted in Figure 3, measured GDP consists of output sold to consumers and investment goods. If the conventional measure of investment in final demand is expanded to include intangibles, then spending on intangibles is no longer treated as an intermediate expenditure, and measured GDP is larger. The rise in output is a first-order impact of capitalizing intangibles in GDP accounting.

In a standard growth-accounting framework, output growth can be decomposed into contributions from capital, from labor, and from total factor productivity growth. What is different in the model that includes intangibles is that there is both an expansion of output, as above, but also—another first-order implication—an expansion of inputs. The contribution of paid-for, commercially valuable knowledge becomes an additional accountable source of output growth, a *direct* contribution if you will. Because intangible capital is only partially appropriable, however, the augmented model also includes a way for intangible capital to explain changes in total factor

productivity. The contribution via total factor productivity is not directly measured, but it reflects the impact of the diffusion, or spread, of innovations embodied in current and past vintages of intangible capital as they are freely copied and adopted across firms and industries in an economy.

As we will discuss later in this paper, this diffusion process, termed "knowledge spillovers" (or increasing returns), is usually modeled as driven by the growth of knowledge itself but may also be affected by institutional factors like the rules concerning patents, trade secrets, and intellectual property, as well as by specific characteristics of the intangible investments themselves.

Why Intangible Capital Is Difficult to Measure

The macroeconomic analysis of intangible capital set out in this paper is grounded in concepts and measures aligned with national accounts. For example, national accounts use investment flows and depreciation rates to derive asset stocks, asset values, and asset incomes. But seeking to apply this approach to a broad category of intangibles is challenging. In this section, we explore several issues.

First, it is often difficult to identify the investment flow, especially when intangible assets are co-produced along with primary products. Second, absent "arm's length" transactions in markets with prices, how can we calculate a price deflator for intangible assets, so that past investments can be expressed in real terms? Third, given that intangible assets lack "substance" (as financial accountants describe this asset class) how should we think of their capital consumption/economic depreciation? Finally, does partial appropriability provide a sufficient conceptual rationale for cumulating and aggregating real flows of intangible investment into capital stocks, as is typically done for tangible assets? This question is relevant for those who question how the competitive advantage of a single firm as reflected in, say, its marketing assets, can create aggregate value for an industry or market in a way that contributes to total factor productivity. These topics are reviewed with reference to "the perpetual inventory method" (PIM), a calculation that assumes depreciation of each asset is geometric and constant across all vintages of the asset and that asset investment flows may be cumulated to obtain measures of real asset stocks.[9]

Investment Flows

Intangible assets may in some cases be acquired via market transactions, like purchases of customer management software systems or of strategic management consulting advice. But more commonly, they are produced within an organization, as in the case of customized software to determine seating in the firm's open office

[9]More specifically, PIM measures the real stock R of individual asset a for a given industry at time t as $R_{(a,t)} = N_{(a,t)} + (1 - \delta_a^R)\ R_{(a,t-1)}$, where $N_{(a,t)}$ is the real investment flow for asset a in the industry. Once each $R_{(a,t)}$ for an industry is obtained, the usual procedures for aggregating over assets and industries are applied.

space or to manage its unique order book. The tendency toward in-house production of intangible capital contrasts with the typical "arm's length" production of most tangible capital. Very few firms make their own tangible assets: for example, UPS does not make its delivery trucks.

A *sum-of-costs* approach is used to estimate investment via in-house production in the macroeconomic data in national accounts. The idea is to imagine a firm, a bank, say, as having a "software factory" or "strategy factory" inside of it and the measurement challenge is to estimate the value of output produced by this hypothetical factory based on factor costs (labor, capital, and intermediates). The linchpin of this approach is identifying the occupations of the workers in the in-house "factory" and estimating their wages and employment from, for example, labor force surveys. From that, the total payments made to all factors used in the in-house production can be estimated. An important assumption in this estimation is the fraction of time spent by the identified workers on the relevant activity. This factor for own-account software investment in the macroeconomic data for many countries is about one half—that is, software developers are assumed to spend one-half of their work time creating new software that is long-lasting. However, this estimate varies within occupational categories, such that software managers are assumed to only spend 5 percent of time on creating long-lived capital.

Could own-account intangible investment be determined more accurately via a survey instrument? Collecting information via a survey instrument is already a proven approach for research and development, which is amenable to data collection via survey because it is well-defined as a business function. European countries gather regular information on firms' expenditures on formal employer-provided training, internal and external, reflecting the fact that training budgets are usually well-defined components of business expenditure. However, own-account investments in software are not well-defined as a separate business expenditure category, nor are the "skunk works" of divisions focused on internal innovation. Surveys of capital expenditures have attempted to collect information on software investments in several OECD countries, including the United States, but results have tended to yield implausibly small figures. Thus, software and databases, and the data series for industrial design, brand, and organizational capital all contain own-account components that are estimated based on the sum-of-costs approach. The series for new financial products consists solely of own-account production.

Asset Price Deflators

An asset price deflator is needed to express past investment in real terms. Because many intangible assets do have a purchased component, a common approach is to use a services output price as an asset price for the deflation of intangibles (for example, Martin 2019). In early empirical work, Corrado, Hulten, and Sichel (2009) used an overall business output price "as a placeholder" in the absence of information on intangible asset prices, noting that this essentially implies that upstream input costs and productivity are little different from downstream (or existing, measured business) sector costs and productivity.

A more sophisticated version of this approach is to identify upstream costs (which may differ substantially in composition from downstream production costs) and apply a productivity adjustment. This approach is in fact used to derive price deflators for business research and development in the US national income and product accounts. The US Bureau of Economic Analysis selected the approach after examining several alternatives (including available service price deflators for the R&D services industry, as discussed in Robbins et al. 2012). The productivity adjustment is a trend derived from the official estimates of nonfarm business sector total factor productivity as published by the US Bureau of Labor Statistics.

Research on hard-to-measure services prices typically does not address intangible asset-producing activities—like R&D labs, marketing teams, engineering design projects—nor are these activities typically viewed as hotbeds of rapid quality change missed by price collectors in assessments of productivity mismeasurement. But more recently, with the digital transformation of economies, the rise of digitally enabled business models, and the increased use of data in business more generally, the nature and efficiency of intangible asset-producing activities arguably have been transformed. This would be manifest in the upstream/downstream model as more rapid total factor productivity growth in the upstream sector, and competitive issues aside, lower prices of intangible asset. Recent developments in intangible asset prices are discussed in Corrado (2021) and analyzed in the context of data, intangibles, and productivity in Corrado et al. (2022).

Economic Depreciation

One might start by asking how knowledge-based intangible assets can even "depreciate": after all, the Pythagorean theorem (and even some Greek buildings!) seems to have lasted for a very long time. But because intangibles are non-rival and returns to investments are not fully appropriable, the value of the investment to the firm or innovator is limited to the returns that the owner/investor can capture. Partial appropriability implies, in stark contrast to the notion that the depreciation of intangible assets must be "slow" because ideas last a very long time, that the value of commercial knowledge declines rather rapidly. This pattern is documented in empirical studies (reviewed in de Rassenfosse and Jaffe 2017; see also Pakes and Schankerman 1984 and Martin 2019) and is supported by survey evidence that asks firms to report the average useful life of their intangible assets (Awano et al. 2010).

Economic depreciation is the reduction in value of an asset as it ages—a price concept that is unobservable and necessary to estimate for any type of capital, tangible or intangible. The definition of economic depreciation showcases the difficulty with textbook explanations of depreciation as physical decay or "wear and tear." Such explanations lose sight of the larger conceptual issue that assets tend to yield less revenue and lose productive value as they age, a loss that reduces value to the firm. All told, then, intangibles do decline in value as firms cease to appropriate benefits because commercialized ideas are replaced by new ones or copied by competitors.

Competitive Advantage and Aggregation

Should investments in marketing assets or brand development, which businesses undertake as a form of competition and to gain a competitive advantage, be conceptually viewed as "capital"? At their root, the question turns on two subsidiary issues. First, do the spending streams for these categories have the longevity that we typically expect of capital? Second, if competing firms both engage in marketing and brand management strategies, would it be more accurate to say that marketing and brand management efforts have some tendency to cancel each other out, rather than the spending by each firm adding up to an overall capitalization value? These topics are discussed in more detail in Corrado (2021), Haskel and Westlake (2019, pp. 49–52), and the paper in this symposium by Bronnenberg, Dubé, and Syverson.

The conceptual basis for treating spending on marketing and brand development as capital is grounded in signaling theory (Milgrom and Roberts 1986), supported by many structural modeling/competition studies, and consistent with the welfare-enhancing effects of product differentiation (Dixit and Stiglitz 1977). The key insight of this broad spectrum of works is that the appropriable revenue stream due to marketing and promotion is determined in general equilibrium via both price and quantity channels. An implication of this view is that product prices are not necessarily higher due to the costs of marketing and promotion. The available empirical evidence also suggests that promotion exhibits important scope economies (for example, it interacts with how a firm chooses to focus its R&D efforts) and that product advertising has, on average, long-lasting informative effects on economic activity in both product markets (as in Rauch 2013) and services industries (as in Kwoka 1984).

In addition, while the original context of much work on intangibles focused on technological innovation via investments in research and development, the analysis of intangibles also has roots in the industrial organization literature, which has focused on the supporting role of marketing in innovation (Hulten 2011). The complementarity between R&D and promotion, both theoretically and empirically, is an established characteristic of globally innovative pharmaceutical firms (Clarkson 1977; Vinod and Rao 2000), as well as other manufactures (for example, Clarkson 1996). In firm-level work on the growth drivers of the software company Microsoft, Hulten (2011) found an important supporting role for marketing in the company's innovation, and a firm-level study of retailers (Crouzet and Eberly 2018) argued that the growing value of brand supported the more efficient practices that spurred the expansion of large retailers in the United States.

In short, the argument that marketing, brand management, and similar activity are only a zero-sum battle breaks down in the presence of innovation and the realities of how modern companies create competitive advantage and differentiate their products. Perhaps a more pertinent question is why macroeconomic practitioners have not been persuaded by the corpus of research on these topics. After all, it is apparent that for marketing assets to have no net impact on aggregate economic activity via consumption as a component of net worth, investments in them must have zero impact on aggregate market capitalization, which would contradict the body of evidence that branding does influence market valuations of firms.

Productivity in Economies with Intangible Capital

We have already demonstrated that measuring intangible capital affects investment/GDP and rates of return. This section focuses on productivity, including remarks on the productivity slowdown and increased role of proprietary data in commercially valuable knowledge. Recent work that has approached measuring and analyzing data as an intangible asset reveals that data capital overlaps almost completely with intangible capital, both conceptually and empirically (Corrado et al. 2022). This change in the composition of intangible capital may have diminished its potential for increasing returns to the extent that the data capital of individual firms is unable to be copied for costless use elsewhere in economies.

To calculate productivity, we use the recently issued EU KLEMS & INTANProd database, which reports productivity estimates including harmonized investment streams for the intangible assets listed in Table 2 for most of Europe, as well as for the United States and Japan.[10] The investment and capital estimates for assets not regularly capitalized in national accounts are developed using methods consistent with national accounts (such as perpetual inventory models): the estimates are not calibrations of a model or developed from data in company financial reports. The methods used to develop the harmonized estimates of intangible investment are documented in Bontadini et al. (2022).[11]

In this section, we report and analyze estimates of total factor productivity that cover ten European countries and the United States from 1998 to 2018. The European aggregate consists of Austria, Germany, Denmark, Spain, Finland, France, Italy, Netherlands, Sweden, and United Kingdom. Future work may bring in more countries—EU KLEMS & INTANProd includes estimates of intangible investment for all 27 EU countries (though histories are short for some). The EU KLEMS & INTANProd data is updated as National Accounts data are released, and so the results here are a snapshot as of March 2022.

Growth Decompositions

The growth accounting reported below is in per hour terms—that is, it decomposes the growth in output per hour for both the European aggregate and the United States. The accounting for the European aggregate is developed at the country-industry level, where industries are aggregated to "market" sector aggregates for each

[10] This update/expansion is funded by the European Commission's Directorate General for Economic and Financial affairs (procurement procedure ECFIN/2020/OP/0001 – Provision of Industry level growth and productivity data with special focus on intangible assets – 2020/S 114-275561).

[11] Available on the EU KLEMS & INTANProd portal at https://euklems-intanprod-llee.luiss.it. Compared with previous estimates for Europe and the United States issued via the INTANInvest database and website (www.intaninvest.net), current figures reflect significant improvements to the own-account components of intangible investment and to intangible asset price deflators. Methods used to develop the current estimates of intangible investment are set out in the appendix to this paper. Regarding deflators for software and tangibles, as in our own previous work, the product quality change component of price deflators for information technology equipment and software is harmonized across countries. The harmonized IT equipment and software deflators are developed and kindly supplied by the OECD.

country and then weighted accordingly to form the European aggregate. Market sector aggregates exclude the public sector and majority-public industries, resulting in coverage across twelve industries that is broadly similar, though not identical, to the nonfarm business sector used for headline productivity statistics in the United States.[12]

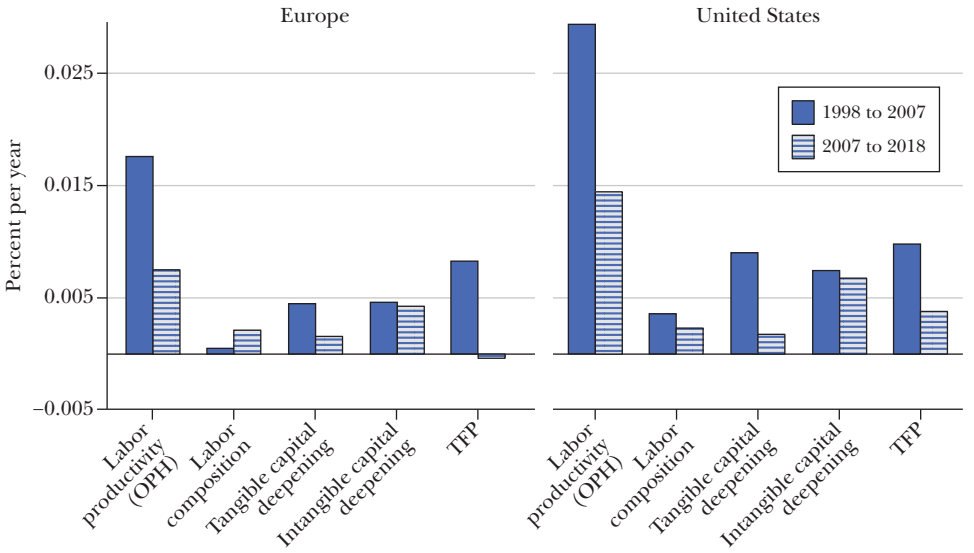
As is well-known, estimates of changes in country-level output per hour reflect both "within" and "between" industry sector effects. The reallocation of labor across sectors is the "between" effect. In lower-income countries, for example, the movement from agriculture to manufacturing is an important source of productivity growth. For high-income countries in recent decades, the main movement across sectors is from manufacturing to services. However, we find that the reallocation of hours across industry sectors has had a negligible impact on broad changes in market sector output per hour in Europe and the United States between 1998 and 2018. When labor productivity growth dropped precipitously in market-dominated industries of both regions with the onset of the global financial recession in 2008, it was almost entirely due to a "within" effect that reached across industries. (By contrast, labor productivity during the pandemic-affected years 2020–2021 is heavily driven by reallocation effects.)

Figure 5 sets out decompositions of industry-aggregated (that is, within-industry aggregates) of labor productivity growth for ten European countries and the United States for the decade leading up to the Great Recession, 1998–2007, and then for 2007–2018. The first pairs of columns for each area shows a substantial drop in labor productivity growth (output per hour) in both areas (-1.0 and -1.5 percentage points, respectively). The last pair of columns reports total factor productivity for each region; they show that the drop in growth of output per hour in Europe is largely accounted for by a substantial slowdown in total factor productivity growth of 0.9 percentage points; similarly, total factor productivity slowed 0.6 percentage points in the United States. The contribution of the second set of bars (labeled labor composition) reflects the per-hour contribution of increases in (employed) human capital, which reflects changes in the proportion of high-skilled/high-wage jobs in industries. Though this effect works in opposite directions in Europe and the United States, its contribution to explaining developments in productivity growth in these regions during the past 20 years is relatively small.

Capital deepening is part of the story of the slowdown in output-per-hour, directly and indirectly. A drop in the contribution of tangible capital deepening directly accounts for nearly one-third of the drop in output-per-hour in Europe and one-half of the drop in the United States. The contribution from the rate at which workers in both regions were equipped with intangible capital edged down only

[12] The market sector aggregates are formed using twelve individual industries that cover ten NACE letter level industry sectors: B (Mining), C (Manufacturing), D and E (Gas, Electricity, and Water), F (Construction), G (Wholesale and retail Trade; repair of motor vehicles), H (Transportation and storage), I (Accommodation and food services), J (Information and Communication activities), K (Finance and insurance activities), M (Professional, scientific, and technical activities), N (Administration and support activities) and R (Arts, entertainment, and recreation). NACE is an international system for industry classification used in Europe; for a concordance to the NAICS system used in North America, see the Bontadini et al. (2022) documentation on the EU KLEMS & INTANProd project portal.

Figure 5
Decompositions of Labor Productivity Growth



Source: Authors' elaboration of EU KLEMS & INTANProd (Bontadini et al. 2022), accessed March 23, 2022.
Note: Decompositions are derived from industry-level data; figures are reported in natural logarithms. See text for composition of the European aggregate. Labor composition and capital deepening columns are labor and capital payments share-weighted per-hour input growth rates. OPH stands for "output per hour." Total factor productivity (TFP) is a residual.

very slightly, however, and thus *directly* explains little of the drop in labor productivity. This finding—which should *not* be interpreted as suggesting that correcting for mismeasurement of intangibles deepens the productivity slowdown puzzle—is discussed further below.

Even though the focus of this article is that national accounts and productivity calculations are missing many intangible assets, the ongoing controversy that official statistics miss major aspects of how consumers benefit from the digital economy cannot be overlooked. For example, the falling cost of consumer digital content delivery—and thus the value that consumers obtain from their paid-for wireless data, internet, and video subscription services—is not well-reflected in GDP. Available research quantifies very fast drops in prices for consumer digital services, especially for mobile data and streaming services, and also increased shares of consumer spending allocated to subscriptions for these services. These are telltale signs that the missed price drops have an *increasing* deflationary impact on consumer price inflation.[13] The missed price drops are in fact estimated to have understated the

[13] The ways in which consumers benefit from free content delivered via their paid-for digital services, like value derived from user-generated content in social media, is a related matter. But however significant, these impacts fall outside the market activity scope of the productivity analysis reported in Figure 5.

deceleration in consumer price change by 0.3 percentage points per year from 2007 to 2018, which when translated to Figure 5, potentially explains one-third to one-half of the estimated drop in growth of total factor productivity. The aggregate estimate is from Byrne and Corrado (2020, 2021), which applies to the United States and covers mobile voice and data, internet access, cable TV, and video streaming. This estimate is consistent with results showing comparably rapid rates of price drops for mobile voice and data in the United Kingdom by Abdirahman, Coyle, Heys, and Stewart (2020) and for music streaming globally by Edquist, Goodridge, and Haskel (2021).

Diffusion of Commercial Knowledge and Increased Productivity Dispersion

The diffusion of commercially valuable knowledge is, logically, a primary determinant of total factor productivity growth according to the upstream/downstream model of Figure 3. The real world is more complex than the basic model, but a connection from intangible capital to productivity growth is a regularity in past productivity data, insofar as cross-country and firm-level econometric work have estimated increasing returns (or knowledge spillovers) to intangible capital. In simple terms, these works imply that a proportional relationship, such that about one-fifth of the growth of intangible capital translates into gains in total factor productivity.[14] The proportional relationship can be used to represent the costless diffusion of commercially valuable knowledge in an economy.

Spillovers are estimated to occur in proportion to the input, not the input-per-hour terms in Figure 5 (the spillovers from a phone network are from the existence of the network, not with the network per hour worked). Intangible capital input did slow in Europe after the financial crisis, from 4.2 percent per year from 1998 to 2007 to 3 percent in the post-crisis period. A spillover effect of one-fifth would predict a total factor productivity slowdown of 0.25 percent in Europe. So, a small part of the total factor productivity slowdown in Europe can be attributed to slower growth in intangible capital; in the United States, the impact is even smaller, less than 0.1 percent.

Another endogenous explanation for the slowdown in measured total factor productivity growth is that the drivers of these increased returns ceased to operate as strongly as they previously had. Why might this change have occurred? One possibility is that the potential for productivity spillovers to intangible investments is determined by an innovation ecosystem, including competition intensity and regulation, intellectual property rights and their enforcement, privacy laws, broadband access, and other factors. It is very difficult, however, to see how the workings of this system could change so seriously and suddenly on both sides of the Atlantic (for some evidence on this point, see Akcigit and Ates 2021).

An alternative possibility is that the composition of knowledge assets directly affects the strength of the diffusion process. Some forms of intangible capital—datasets,

[14] This underlying estimates here refer to the aggregate implications of estimates for R&D spillovers reported by Griliches (1992, 1994) for manufacturing and the similar estimates for non-R&D intangibles (especially, the industrial design, employer-provided training, and organizational capital components) by Corrado, Haskel, and Jona-Lasinio (2017).

certain formulas, and software code—tend to be regarded as trade secrets, intentionally undisclosed and difficult to replicate. The digital economy has boosted the share of investment in these forms, which arguably weakens mechanisms that generate increasing returns to intangible capital. As intangible capital has become, in effect, data capital, there also has been an increase in dispersion of firm-level productivities *within* industry groups attributed, at least in part, to increased investments in economic competencies by market services industries. This pattern was documented globally in Andrews, Criscuolo, and Gal (2016), who characterized the development as a worrisome decline in the global diffusion of new ideas and technologies since 2000. The growing relative importance of intangible assets was identified as a mechanism behind increased firm-level productivity dispersion in follow-on work (Corrado et al. 2021). This changed composition of intangible investment then may also have led to scale economies within *certain* firms, like data agglomeration effects in digitally enabled firms, that tended to reduce competition in those markets.

In the intangible capital framework, the maximum impact of these developments on market sector productivity is as follows: With post-2007 growth of intangibles averaging 3 percent per year in the European countries and about 3½ percent per year in the United States, and applying the approximation that one-fifth of this growth translates into a change in total factor productivity, a *complete* cessation of the diffusion mechanism would shave more than ½ percentage point per year off measured total factor productivity growth in these regions. Productivity growth via the costless replication of commercial knowledge is of course highly unlikely to have ceased entirely, and this brief analysis does not rule out other possible culprits behind the productivity slowdown. But the increased use of data and increasing overlap between data capital and intangible capital is an important development that is likely having an impact on productivity growth in modern economies.

Final Remarks

The framework for intangible capital presented here builds bridges between GDP measurement, growth accounting, and modern growth theory: because intangibles are also nonrival, productivity narratives using the intangible capital framework naturally embrace endogenous factors that modern growth theory emphasizes. In its focus on the partial appropriability of investments in innovation, the intangibles framework provides economists with a bridge to discussions of methods of business innovation in the management literature. Several key topics related to intangible capital have received no mention or only a very light touch here, such as how digital technologies like cloud computing and artificial intelligence affect productivity and how data assets are captured in the intangible capital framework studied elsewhere (Corrado, Haskel, and Jona-Lasinio 2021; Corrado et al. 2022), as well as the policy-related dimensions of intangible capital reviewed in Haskel and Westlake (2022).

The trendlines suggest that the intangible economy is only becoming more important. Policymakers, along with policy and business analysts, are already putting

intangible capital into economic frameworks used for analysis: some examples with which we are familiar include central banks, the OECD, European Commission, Italian G20 Presidency, and business-oriented research organizations such as the Conference Board, McKinsey Global Institute, and NESTA (UK). As modern economies become more "knowledge-intensive," we believe that economic researchers should seek to include the full complement of intangibles in investment, profits, and productivity data. Continued movements in this direction by statistical agencies and data compilers will make business data much more representative of the intangible world around us.

References

Aaker, David A. 1991. *Managing Brand Equity*. New York: Free Press.

Akcigit, Ufuk, and Sina T. Ates. 2021. "Ten Facts on Declining Business Dynamism and Lessons from Endogenous Growth Theory." *American Economic Journal: Macroeconomics* 13 (1): 257–98."

Abdirahman, Mohamed, Diane Coyle, Richard Heys, and Will Stewart. 2020. "A Comparison of Deflators for Telecommunications Services Output." *Economie et Statistique* 517-518-519: 103–22.

Andrews, Dan, Chiara Criscuolo, and Peter Gal. 2016. *The Global Productivity Slowdown, Technology Divergence and Public Policy: A Firm Level Perspective*. Paris: OECD.

Awano, Gaganan, Mark Franklin, Jonathan Haskel, and Zafeira Kastrinaki 2010. "Measuring Investment in Intangible Assets in the UK: Results from a New Survey." *Economic and Labour Market Review* 4 (7): 66–71.

Bontadini, Filippo, Carol, Corrado Jonathan Haskel, Massimiliano Iommi, and Cecilia Jona-Lasinio. 2022. *The EUKLEMS & INTANProd Database: Methods and Data Descriptions*. Rome: LUISS.

Brynjolfsson, Erik, and Lorin M. Hitt. 2000. "Beyond Computation: Information Technology, Organizational Transformation and Business Performance." *Journal of Economic Perspectives* 14 (4): 23–48.

Brynjolfsson, Erik, Lorin M. Hitt, and Shinkyu Yang. 2002. "Intangible Assets: Computers and Organizational Capital." *Brookings Papers on Economic Activity* 1 (1): 137–99.

Byrne, David, and Carol Corrado. 2020. "The Increasing Deflationary Influence of Consumer Digital Access Services." *Economic Letters* 196 (109447): 1–4.

Byrne, David, and Carol Corrado. 2021. "Accounting for Innovations in Consumer Digital Services: IT Still Matters." In *Measuring and Accounting for Innovation in the Twenty-First Century*, Vol. 78, *Studies in Income and Wealth*, edited by Carol Corrado, Jonathan Haskel, Javier Miranda, and Daniel Sichel, 471–517. Chicago: University of Chicago Press.

Clarkson, Kenneth W. 1977. *Intangible Capital and Rates of Return: Effects of Research and Promotion on Profitability*. Vol. 138, *American Enterprise Institute Studies in Economic Policy*. Washington, DC: American Enterprise Institute for Public Policy Research.

Clarkson, Kenneth W. 1996. "The Effects of Research and Promotion on Rates of Return." In *Competitive Strategies in the Pharmaceutical Industry*, edited by Robert B. Helms, 238–68. Washington, DC: AEI Press.

Corrado, Carol. Forthcoming. "Measuring Intangible Capital: Implications for Growth and Productivity." Brookings Hutchins Center Working Paper.

Corrado, Carol, Chiara Criscuolo, Jonathan Haskel, Alexander Himbert, and Cecilia Jona-Lasinio. 2021. "New Evidence on Intangibles, Diffusion and Productivity." OECD Science, Technology and Industry Working Paper 2021/10.

Corrado, Carol, Peter Goodridge, and Jonathan Haskel. 2011. "Constructing a Price Deflator for R&D: Calculating the Price of Knowledge Investments as a Residual." The Conference Board Economics Program Working Paper 11–03.

Corrado, Carol, Jonathan Haskel, Massimiliano Iommi, Cecilia Jona-Lasinio, and Fillipo Bontadini. 2022. "Data, Intangibles and Productivity." Paper presented at the NBER/CRIW Conference on Technology, Productivity, and Economic Growth, Washington, DC, March 17.

Corrado, Carol, Jonathan Haskel, and Cecilia Jona-Lasinio. 2016. "Intangibles, ICT and Industry Productivity Growth: Evidence from the EU." In *The World Economy: Growth or Stagnation?*, edited by Dale W. Jorgenson, Kyoji Fukao, and Marcel P. Timmer, 319–346. Cambridge, UK: Cambridge University Press.

Corrado, Carol, Jonathan Haskel, and Cecilia Jona-Lasinio. 2017. "Knowledge Spillovers, ICT, and Productivity Growth." *Oxford Bulletin of Economics and Statistics* 79 (4): 592–618.

Corrado, Carol, Jonathan Haskel, and Cecilia Jona-Lasinio. 2021. "Artificial Intelligence and Productivity: An Intangible Assets Approach." *Oxford Review of Economic Policy* 37 (3): 435–58.

Corrado, Carol, Charles R. Hulten, and Daniel Sichel. 2005. "Measuring Capital and Technology: An Expanded Framework." In *Measuring Capital in the New Economy*, Vol. 66, Studies in Income and Wealth, edited by Carol Corrado, John Haltiwanger, and Daniel Sichel, 11–46. Chicago: NBER, University of Chicago Press.

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

Corrado, Carol, and Charles Hulten. 2010. "How Do You Measure a 'Technological Revolution'?" *American Economic Review* 100 (2): 99–104.

CPA Ontario. 2021. *You Can't Touch This: The Intangible Assets Debate.* Toronto: Chartered Professional Accountants of Ontario.

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

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

Cuthbertson, Richard, Peder Inge Furseth, and Stephen J. Ezell 2015. "Apple and Nokia: The Transformation from Products to Services." In *Innovating in a Service-Driven Economy: Insights, Application, and Practice*, 111–29. London: Palgrave Macmillan.

de Rassenfosse, Gaétan, and Adam B. Jaffe. 2017. "Econometric Evidence on the R&D Depreciation Rate." NBER Working Paper 23072.

Dixit, Avinash K., and Joseph E Stiglitz. 1977. "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67 (3): 297-308.

Döttling, Robin, and Lev Ratnovski. 2020. "Monetary Policy and Intangible Investment." IMF Working Paper 2020/160.

Edquist, Harald, Peter Goodridge, and Jonathan Haskel. 2021. "The Economic Impact of Streaming beyond GDP. *Applied Economics Letters* 29 (5): 403–8.

Erdem, Tülin, and Michael P. Keane. 1996. "Decision-making Under Uncertainty: Capturing Dynamic Brand Choice Process in Turbulent Consumer Goods Markets." *Marketing Science* 15 (1): 1–20.

Erdem, Tülin, Michael P. Keane, and Baohong Sun. 2008. "A Dynamic Model of Brand Choice When Price and Advertising Signal Product Quality." *Marketing Science* 27 (6): 949–1136.

Farquhar, Peter H. 1989. "Managing Brand Equity." *Marketing Research* 1 (3): 24–33.

Greenstein, Shane. 2016. "What Does a Skunk Works Do?" *IEEE Micro* 36 (2): 70–1.

Griliches, Zvi. 1973. "Research Expenditures and Growth Accounting." In *Science and Technology in Economic Growth*, edited by B.R. Williams, 59–95. London: Palgrave MacMillan.

Griliches, Zvi. 1979. "Issues in Assessing the Contribution of Research and Development to Productivity Growth." *Bell Journal of Economic*s 10 (1): 92–116.

Griliches, Zvi, 1986. "Productivity, R&D, and the Basic Research at the Firm Level in the 1970's." American Economic Review 76 (1): 141–154.

Griliches, Zvi. 1992. "The Search for R&D Spillovers." *Scandinavian Journal of Economics* 94: S29–47.

Griliches, Zvi. 1994. "Productivity, R&D, and the Data Constraint." *American Economic Review* 84 (1):1–23.

Groysberg, Boris, Andrew N. McLean, and Nitin Nohria. 2006. "Are Leaders Portable?" *Harvard Business Review* 84 (5): 92–100.

Haskel, Jonathan, and Stian Westlake. 2018. *Capitalism without Capital: The Rise of the Intangible Economy.* Princeton: Princeton University Press.

Haskel, Jonathan, and Stian Westlake. 2022. *Restarting the Future: How to fix the intangible economy.* Princeton: Princeton University Press.

Hulten, Charles R. 2010. "Decoding Microsoft: Intangible Capital as a Source of Company Growth." NBER Working Paper 15799.

Hulten, Charles R. 2011. "How did Microsoft Become 'Microsoft'? Intangible Capital and the Endogenous Growth of the Firm." Unpublished.

Jones, Charles I. 2019. "Paul Romer: Ideas, Nonrivalry, and Endogenous Growth." *Scandinavian Journal of Economics* 121 (3): 859–83.

Jorgenson, Dale W. 1963. "Capital Theory and Investment Behavior." *American Economic Review* 53 (2): 247–59.

Kwoka, John. 1984. "Advertising and the Price and Quality of Optometric Services." *American Economic Review* 74 (1): 211–6.

Lev, Baruch. 2001. *Intangibles: Management, Measurement, and Reporting*. Washington, DC: Brookings Institution Press.

Lev, Baruch, and Suresh R. Radhakrishnan. 2005. "The Valuation of Organization Capital." In *Measuring Capital in the New Economy*, edited by Carol Corrado, John Haltiwanger, and Daniel Sichel, 73–110. Chicago: NBER, University of Chicago Press.

Martin, Josh. 2019. "Measuring the Other Half: New Measures of Intangible Investment from the ONS." *National Institute Economic Review* 249 (1): R17–29.

Milgrom, Paul, and John Roberts. 1990. "Bargaining Costs, Influence Costs, and the Organization of Economic Activity." In *Perspectives on Positive Political Economy*, edited by James Alt and Kenneth A. Shepsle, 57–89. Cambridge, UK: Cambridge University Press.

Milgrom, Paul, and John Roberts. 1986. "Price and Advertising Signals of Product Quality." *Journal of Political Economy* 94 (4): 796–821.

Moris, Francisco. 2018. "Definitions of Research and Development: An Annotated Compilation of Official Sources." Alexandria, VA: National Science Foundation.

Nakamura, Leonard. 1999. "Intangibles: What Put the *New* in the New Economy?" *Federal Reserve Bank of Philadelphia Business Review* (July/August): 3–16.

Nakamura, Leonard. 2001. "What is the U.S. Gross Investment in Intangibles? (At Least) One Trillion Dollars a Year!" Federal Reserve Bank of Philadelphia Working Paper 01-15.

OECD. 2013. *Supporting Investment in Knowledge Capital, Growth and Innovation*. Paris: OECD Publishing.

OECD. 2015. *Frascati Manual 2015—Guidelines for Collecting and Reporting Data on Research and Experimental Development*. Paris: OECD Publishing.

OECD. 2017. *OECD Science, Technology and Industry Scoreboard 2017: The Digital Transformation*. Paris: OECD Publishing.

Pahwa, Nitish. 2021. "How the Heck Is Peloton the Best-Paying Music Streaming Service?" *Slate*, July 12, https://slate.com/culture/2021/07/peloton-music-royalties-spotify-apple-music.html.

Pakes, Ariel, and Marcus Schankerman. 1984. "The Rate of Obsolescence of Patents, Research Gestation Lags, and the Private Rate of Return to Research Resources." In *R&D, Patents, and Productivity*, edited by Zvi Griliches, 73–88. Chicago: University of Chicago Press.

Piekkola, Hannu. 2016. "Intangible Investment and Market Valuation." *Review of Income and Wealth* 62 (1): 28–51.

Robbins, Carol, Olympia Belay, Matthew Donahoe, and Jennifer Lee. 2012 "Industry-level Output Price Indexes for R&D: An Input-cost Approach with R&D Productivity Adjustment." Unpublished.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5): S71–102.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2019. "Capitalists in the Twenty-First Century." *Quarterly Journal of Economics* 134 (4): 1675–745.

Vinod, H.D., and P.M. Rao. 2000. "R&D and Promotion in Pharmaceuticals: A Conceptual Framework and Empirical Exploration." *Journal of Marketing Theory and Practice* 8 (4): 10–20.

Young, Alison. 1998. *Towards an Interim Statistical Framework: Selecting the Core Components of Intangible Investment*. Paris: OECD.

Zhang, Daniel, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, et al. 2021. *The AI Index 2021 Annual Report*. Stanford: Stanford University Human-Centered AI Institute.

Zwick, Thomas. 2007. "Apprenticeship Training in Germany? Investment or Productivity Driven?" ZEW Discussion Paper 07-023.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 29–52

The Economics of Intangible Capital

Nicolas Crouzet, Janice C. Eberly,
Andrea L. Eisfeldt, and Dimitris Papanikolaou

Intangible capital is generally defined by what it lacks—that is, as productive capital that lacks a physical presence. Familiar and important examples include patents, software and databases, trademarks, customer lists, franchise agreements, and organization capital and firm-specific human capital.

In contrast, we will focus on the properties that affirmatively characterize intangibles. Fundamentally, since intangibles lack a physical presence, they require a *storage medium*. The medium can be a piece of physical capital, like a computer (for software), or a document (for a patent or a design), or a person (for a method or an innovation). This need for a storage medium has important implications for the two properties that we emphasize throughout the paper. First, intangibles can be used simultaneously in production in different locations and processes, which implies some degree of *non-rivalry in use*. Because the same intangible can be simultaneously stored (copied) in multiple places and used simultaneously in production, intangibles allow for economies of scale and scope. However, it can also be difficult to establish and enforce exclusive property rights to an intangible; unlike a physical piece of capital, an intangible can be readily copied or imitated, simply by copying

■ *Nicolas Crouzet is Associate Professor of Finance, Northwestern Kellogg School of Management, Evanston, Illinois. Janice C. Eberly is Professor of Finance, Northwestern Kellogg School of Management, Evanston, Illinois. Andrea L. Eisfeldt is Professor of Finance, Anderson School of Management, Los Angeles, California. Dimitris Papanikolaou is Professor of Finance, Northwestern Kellogg School of Management, Evanston, Illinois. Eberly, Eisfeldt, and Papanikolaou are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are n-crouzet@kellogg.northwestern.edu, eberly@kellogg.northwestern.edu, andrea.eisfeldt@anderson.ucla.edu, and d-papanikolaou@kellogg.northwestern.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.29.

software or by learning information, for example. We describe this property as *limited excludability*. Limited excludability makes it difficult to establish an intangible as an asset with enforceable property rights.

We begin by discussing these properties and their implications. We point out that the extent to which these properties generate a valuable intangible asset—which motivates investment—depends on the properties of the storage technology, and the resulting non-rivalry and excludability, and the institutions that enforce property rights.

Figure 1 plots the relative value of internally generated intangible assets to tangible capital. Two things are of note in the figure. First, intangible assets grew much faster than tangible assets throughout the 1990s and early 2000s. Second, the faster growth in intangibles appears to have ceased around 2005 (at least for public firms). Regardless of the current growth rate, it is clear that intangibles represent a very large fraction of corporate capital.

Existing research has largely defined intangibles as a variant within the traditional physical capital framework: for example, the investment process for intangible capital is more uncertain; intangibles depreciate slower (or faster, or more randomly) than physical capital; intangibles have a different relative price; and so on. Intangibles then often amount to "missing" or "mismeasured" capital. Measuring intangible capital is difficult and does tend to exclude significant components (for instance, intangibles stored in employees). Starting from a more affirmative description of intangibles bypasses these boundaries and leads to novel implications for the theory and measurement of intangibles. We discuss a model for capturing the economic implications of these properties and consider its implications. We show how this approach can shed light on some important recent macroeconomic and financial trends, including declining measured productivity growth, growing inequality, rising market power, rising valuations, and declining tangible investment rates.

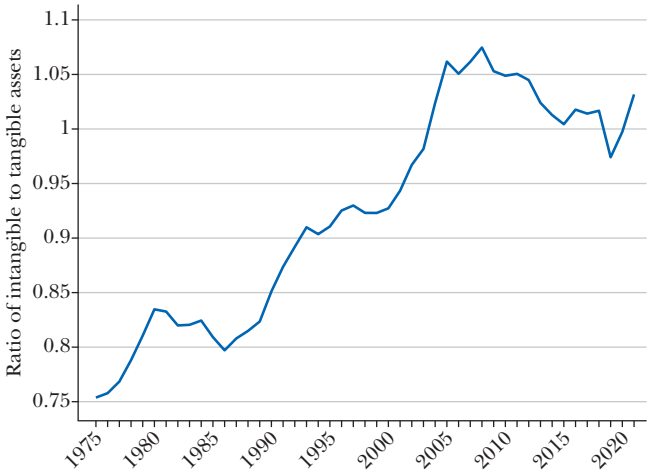
Characteristics of Intangibles as Assets

Two Fundamental Properties

Intangibles are capital, like machinery or structures, in the sense that creating them requires foregoing consumption today (investment) to achieve more output in the future. Unlike machinery or structures, however, intangibles lack a physical presence. At an abstract level, intangibles consist of information. As information without a physical form, intangibles must be stored in some medium in order to be used in production. The need to store intangibles creates their two fundamental properties, which we will call non-rivalry in use and limited excludability.

Storage of intangible assets may be done using different media: speech, writing, drawing, software, recordings, or other technologies. For instance, an algorithm is stored using code or software; a logo is stored using drawings; a managerial process is stored in a team of workers or in written instructions specifying the operational rules of a firm. Although intangibles are sometimes conflated with their storage

Figure 1
Ratio of Intangible Assets to Tangible Assets for US Public Firms, 1975–2021



Source: Authors' calculations (Crouzet et al. 2022a).
Note: Intangible assets are constructed by applying the perpetual inventory method to 30 percent of firms' Selling and General Administrative expenses following Eisfeldt and Papanikolaou (2013, 2014) and Eisfeldt, Kim, and Papanikolaou (2022). Tangible assets are property, plant, and equipment.

medium, the two are distinct. The value of a book is not the paper and cover, but rather the information it holds. Algorithms and data can be written down on paper or encoded in software or databases, but the value derives from the information, not the medium that encodes it.

Given a storage medium, the *non-rivalry in use* property arises because the same intangible can be stored simultaneously in multiple places. By duplicating the storage medium, the same intangible can be used as an input in production across *multiple* goods or services at the *same* time. For instance, the same algorithm can be copied (stored in multiple places) and used in multiple simultaneous instances to produce, say, search results. The same design for a logo can be drawn and then copied and used in multiple simultaneous instances to brand clothing products. Managerial processes can be used in multiple simultaneous instances in different parts of the same organization, or across firms around the world. We specify "non-rivalry in use," as opposed to simply "non-rivalry," to stress that intangibles are production inputs. By contrast, the public economics literature commonly uses the expression "non-rival" to refer to consumption goods, rather than to production inputs. The degree of non-rivalry in use depends on the technology underlying the storage medium: an algorithm stored in teams of workers, for example, may be less efficient to use across locations than one that is stored and deployed in software.

Similarly, even if an intangible is stored in a particular medium, it can be difficult to claim and enforce property rights to the surplus it might create. We refer to

this second property of intangibles as *limited excludability*. An extreme example of intangibles with limited excludability are public goods that can be used as capital inputs, such as an open-source operating system, or a method for making fire. These inputs are non-rival in use within the firm, and also it is not possible to exclude other firms from using them. Other examples of intangible capital—such as patents—offer more property rights protection to their owners. In what follows, we will use "intangible assets" to describe intangible capital inputs whose value can be captured and privately appropriated. The extent to which this is possible depends *both* on the technological features of the storage medium and on the institutional environment.

Technology determines how intangibles are stored. Prior to the development of writing (including images), intangible knowledge was passed down from one person to another through speech. Writing allowed for intangibles to be stored independently of individuals. As technology has progressed, the scope of which intangibles can be stored has expanded. For instance, digital media can store larger amounts of information than writing, allowing for storage of complex intangibles such as genomic sequences or consumer databases. Recording a lecture is a more comprehensive form of storage than distributing the notes for that lecture. These technological advances in storage technologies may help to explain the rapid rise in measured intangibles since the 1990s.

Institutions, both informal and formal, create extrinsic value from stored intangible assets by enforcing excludability, which limits the set of agents with the right to use the asset and capture its value. For instance, limits on the disclosure of ideas, such as trade secrets, create excludability. Excludability is often formalized and enforced through the legal system, including the patent system, copyright enforcement, and non-compete clauses.

There is feedback between technology and institutions. Institutions affect the incentives to store intangibles in different media. For example, the fact that software can be copied will undermine its value, unless intellectual property rights are enforced by institutions. Conversely, the degree of legal protection depends on the storage technology. Reliable storage makes it easier to identify and enforce legal protections. Moreover, as the technology to store intangibles evolves, it can displace the value of intangibles stored in now obsolete technologies: for example, software has replaced many of the human resource functions previously done by trained labor. Intangibles that are harder to codify, such as higher-level management practices, can be harder to imitate—except by hiring away key labor inputs. Differences in displacement risk for intangibles stored in labor inputs may have played a role in trends in income and wealth inequality.

In Table 1, we list some common examples of intangible capital. All have an element of non-rivalry in the sense that underlying information or instructions or contract provisions that make up the intangible capital can be used repeatedly in different times and places, though sometimes with imperfect resolution or reproducibility. In addition, these types of intangible capital vary in how they are stored and how property rights are generated. For example, property-rights enforcement may be centralized (say, via the US Patent and Trademark Office), or bilateral, using contract law.

Table 1
Examples of Intangible Assets: Storage Technology and Property-Rights Institution

	Storage Medium	*Property-Rights Institution*
Patents and blueprints	Patent application	Patent system
Software and databases	Computers	Copyright system
Video and audio material	Audiovisual media	Copyright system
Franchise agreements	Codified rules in contract	Contract enforcement
Consumer lists and purchase agreements	Digital media, contracts, or within employees	Contract enforcement
Organization capital	Key employee talent, manuals	Non-compete clauses, trade secrets
Brands	Consumers, trademark media	Trademark system

Note: A subset of these are drawn from IAS 38, which lays out the criteria for recognizing and measuring intangible assets according to the International Financial Reporting Standards (IFRS) Foundation, available at https://www.iasplus.com/en/standards/ias/ias38.

Parallels with Physical Capital

Intangible capital shares several properties with physical capital: 1) it is an accumulated factor; 2) it depreciates; and 3) it can be firm-specific to varying degrees. We briefly discuss these parallels.

Like physical capital, intangibles require investment, which is commonly observed in functions like research and development, marketing, or human capital and skill accumulation. The mapping from investment to accumulated capital may be less certain for intangible than for physical capital—which is one reason that conventional accounting has historically not capitalized research and development in the same way as physical capital expenditures. Of course, the mapping for physical capital may be less certain than is typically assumed. The measured physical capital stock is not a census of machines—it is an accumulation of investment. Historical expenditure on fiber optic cables, for example, did not accumulate in a simple way to the current value nor productive use of the current stock of fiber optic cable.

For physical capital, depreciation can be caused by wear and tear and by obsolescence. Intangibles do not suffer wear and tear, only obsolescence. Obsolescence of intangible capital can result from several causes: lack of continued investment (maintenance) in intangibles; the arrival of new/or better vintages of intangibles; or other reasons exogenous to the firm. For instance, brand value may be forgotten if marketing expenses are not kept up; management and production processes may become obsolete as new methods appear; knowledge can be lost when employees depart; and data that is not up-to-date becomes less useful. For intangible capital, reversing or slowing the extent of depreciation due to obsolescence requires investment that involves innovation and whose outcome may be more uncertain than replacement of physical capital. In addition, similar to the way in which physical capital may be destroyed as a result of a natural disaster, intangibles can be destroyed by other disasters: corporate scandals; violation of intellectual property

laws by private actors or expropriation by governments; employees with key skills leaving the organization; changes in laws; or shifts in consumer tastes (for example, when a sports team changes its brand name in response to shifting cultural norms). The forces driving depreciation of intangibles can lead to large and abrupt negative shocks in the form of rare disasters to the accumulation of intangible capital.

Finally, the two key properties of intangibles highlighted above—non-rivalry in use, and limited excludability—can also be thought of in the context of physical capital. Physical assets are, by definition, rival in use: a particular truck cannot produce transportation services across different routes at the same time; the same mill cannot produce steel pipes in different locations at the same time. Additionally, property and control rights are generally easier to assign to physical assets. Trucks must be titled, and the title identifies the owner. Ownership of the steel mill, while it might be shared, is formalized through contracts, and disputes regarding control generally have legal remedies.

Production with Intangible Capital

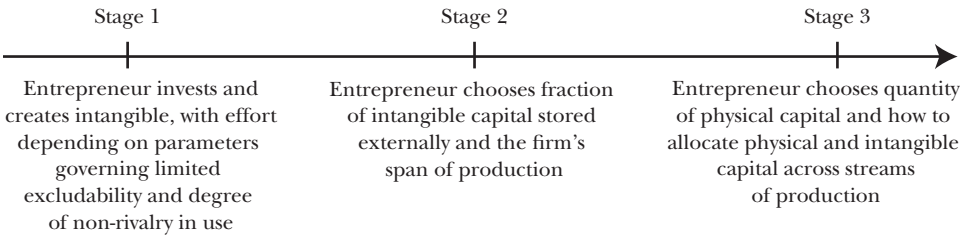
In this section, we describe a model that links the ideas discussed in the previous section—storability, non-rivalry, and limited excludability—and derive their implications for production and investment. A full algebraic presentation of the model is available in the online Appendix with this article at the *JEP* website.

A Model of Production with Intangibles

We focus on a single firm which operates for a single period and makes operating and investment choices to maximize the terminal value of profits. Although we use the word "entrepreneur" to describe the owner of the firm (and the intangible asset), the term is meant more broadly to encapsulate all parties that participate in the creation, dissemination, and use of intangible assets. Examples include all skilled personnel who are responsible for the creation of new inventions or business ideas, entrepreneurs, managers, and startup employees. The model includes two types of capital, physical and intangible. Both types of capital can be deployed across multiple production streams, which can be thought of as different product lines, physical locations, or market segments. The number of production streams determines the span of the firm. To highlight the role of intangibles, we minimize the role of physical capital: in this model, it can be rented at a constant user cost. By contrast, intangibles in the model need to be stored but may be non-rival in use. Moreover, limited excludability will limit the scope of deployment of intangible capital.

For ease of exposition, we split the model into three stages shown in Figure 2: decision about the level of intangible investment, choice of the span of the firm, and the allocation of intangibles and physical capital across production streams. In building intuition about the model, it is convenient to discuss these choices in reverse order.

Figure 2
Timeline of the Model



Source: Crouzet et al. (2022b).

Stage 3: Choice of Physical Capital and Production

In the production stage, the firm's intangible capital and its span of production are taken as given. Thus, the firm chooses the amount of physical capital and the allocation of physical and intangible capital to each stream of production. Again, the firm can rent whatever total stock of physical capital it wants to use. Within each stream, production uses the two inputs, intangible capital and physical capital, with constant returns to scale. In the deployment of physical capital, the same unit (say, a machine) cannot be simultaneously used in multiple production streams. For the profit-maximizing firm, the marginal revenue product of applying physical capital across each production stream will be the same.

In principle, intangibles are non-rival in use, and hence the same intangible can be used simultaneously in different production streams. But in practice, intangible capital need not be completely non-rival within a firm and across production streams. Instead, there can be partial non-rivalry. In our model, the degree of non-rivalry in use for intangible capital within the firm can be thought of as a parameter ranging from 0 to 1. At one extreme, intangibles are rival within the firm. In this case, just like physical capital, using an intangible in a production stream precludes its use in another stream. A luxury brand cannot be used to market household cleaning products, and data about luxury spending can't be used to plan cleaning product inventories. At the other extreme, the same intangible can be used in every production stream—that is, the firm could use its entire stock of intangible capital in each of the production streams. Payroll software can be used across the entire firm, as can a healthy corporate culture. In that case, investing in intangibles (in Stages 1 or 2) may generate economies of scale or scope.

The degree of non-rivalry in use may only be partial—that is, the parameter mentioned above may fall strictly between 0 and 1. This could arise as a result of imperfections in the storage technology. For instance, when intangible capital is stored within key employees, it may be difficult to communicate knowledge perfectly across different parts of the firm—information is often lost in translation. Similarly, software and brands may need to be customized to fit different production locations or product lines.

It is useful to think about partial non-rivalry in use of intangible capital within the firm in terms of the marginal rate of substitution between intangible capital allocated in any two production streams. Remember that each production stream is constant-returns-to-scale in this model, but some production streams can have a higher relative use of intangible capital than others. For a given total stock of intangibles, when there is perfect non-rivalry in use, increasing the intangible asset in one production stream does not require reducing its use in any other production stream. When non-rivalry is partial, increasing the use of intangible capital in one production stream does require reducing its use in another production stream—but less than one-for-one. In other words, increasing the intangible input in one production stream is not entirely costless, but it does not necessarily eliminate its availability for other production streams, either.

A firm will make its Stage 1 and Stage 2 decisions—about intangible capital and the span of the firm across production streams—with the knowledge that at Stage 3 it will be able to rent physical capital as needed. Thus, when thinking about the quantity of intangible capital, the firm will also consider the span of production streams over which that intangible capital could be applied, along with the extent of non-rivalry in use of intangibles across these production streams. In this sense, the quantity of intangible capital and the span of the firm across production streams are complements. In other words, a higher span of production processes further increases the returns to intangible investment. We next discuss how the firm chooses firm span (at Stage 2) and intangible investment (at Stage 1).

Stage 2: Storage Choice and Costs of Expanding Firm Span

At this stage, the firm will take as given the level of intangible capital. In the absence of any cost of increasing firm span, the intangible would be allocated to every possible production stream in order to take advantage of non-rivalry. This section introduces a tradeoff between the value generated by increasing the span of the firm and potential costs of doing so. This cost is not a physical cost, but rather arises endogenously from the limited excludability of intangibles.

We can think of an entrepreneur initially facing a storage choice: whether to retain the intangible as closely held or to codify the intangible and store it in an external medium, such as general labor or capital. This storage decision involves an important tradeoff. By storing more of the intangible externally, an entrepreneur or manager can increase the firm's span of production and better exploit the non-rival nature of intangibles. On the other hand, by codifying the intangible and storing it externally, the entrepreneur reduces the cash flows, control rights, and rents that can be obtained from the intangible. In other words, codifying intangibles increases the size of the firm but reduces the share that is appropriable solely by the entrepreneur. More generally, the choice of storage medium could also affect the degree of non-rivalry. We discuss two specific examples in which greater span undermines excludability: imitation and incomplete contracts.

The issue of imitation arises because as a firm adds new production streams it enables the use of those intangibles across more production facilities, markets,

or geographies. More workers and more consumers are exposed to the firm's intangibles, and so competitors may find it easier to replicate the intangible and to appropriate some of the returns it generates. This effect could vary with the type of intangible and its storage technology. For instance, the imitation problem will tend to be higher for intangibles that are easy to store in capital versus labor (and thus easier to copy) and those that are not well-protected by specific legal institutions, like software or research ideas. This potential for imitation limits the firm's ability to exclude others from appropriating the benefits of its intangibles—and more so as the firm's span rises. This generates a cost of increasing firm span that balances the benefits of non-rivalry.

A cost due to incomplete contracts arises because, if the entrepreneur keeps the intangible closely held, outside investors are subject to potential hold-up problems or information asymmetries. As a result, for the intangible to serve as a basis for external financing, it must first be stored and codified outside of the entrepreneur, so that future cash flow or control rights can be assigned to the firm's investors. Once stored, returns to the intangible can then partially accrue to outside investors as the return on their investment in the firm. The costs of external finance will tend to be higher if the nature of the storage technology for the intangible makes pledgeability difficult. Intangibles stored in key talent, organization capital, and technical advances that are difficult for outside parties to evaluate (so that only the entrepreneur can assess its true value) present challenges to pledgeability. In turn, these challenges to pledgeability mean that the entrepreneur must give up a larger share of the intangible in order to obtain financing. Note that improvements to storage technologies can imply that the creators of intangibles accrue substantial value if the span of those intangibles increases substantially as external storage increases.

Under either approach to limited excludability (imitation or incomplete contracts), the entrepreneur loses the ability to claim fully the benefits of the intangible; this cost is captured by a *limited excludability parameter*, which can be thought of as ranging from 0 to 1. This parameter captures the decline in the share of the total surplus generated by the intangible that actually accrues to the entrepreneur as span rises, owing to the inability to exclude outside agents from claiming the benefits of the intangible. In the case of imitation, this may generate a gap between the private and social choice of span, where the firm may choose a lower span than is socially optimal, by ignoring the potential external benefits of spillovers. In the case of incomplete contracts, the entrepreneur who only controls a fraction of the firm's intangibles will also choose span based on how much of their value she can capture, rather than on the full value of intangibles. Even if external financing brings in funding, losses in translation from codification or mispricing reduces the value of the intangible to the entrepreneur.

The limited excludability parameter can be thought of as capturing limits to property rights or information asymmetries. In the case in which the cost of storing intangibles comes from imitation, stronger and more well-defined property rights can lead to a larger increase in span per unit of codified intangibles. For the case in which codification improves pledgeability and external financing opportunities,

better property rights or more perfect information improves the tradeoff between the increase in span and the rents captured by the creator of the intangible.

The fact that increasing the span of the firm reduces excludability creates a tradeoff relative to the previous subsection. There, increasing span could generate returns to scale or scope—because of the non-rival use of intangibles—which is now constrained by limited excludability. How does the limited excludability parameter interact with the degree of non-rivalry in use parameter discussed in the previous subsection? If an entrepreneur faces a high degree of non-rivalry and also a high level of excludability, there will be an incentive to adopt an intangible-intensive production technology and to operate at high scale. After all, the span of a firm tends to increase with the degree of non-rivalry. However, the entrepreneur will only value the associated intangible asset to the extent that the benefits can be claimed and property rights are strong. If non-rivalry is high but excludability is low, the entrepreneur might instead steer away from investing in intangible capital, and instead pick a technology that emphasizes physical capital inputs and focuses on a single production stream. When it comes to investment in intangible capital, *non-rivalry and excludability are complements.*

Finally, this framework illustrates that an entrepreneur's scale choices may be socially inefficient. If excludability is low, whether for imitation or incomplete contract reasons, the entrepreneur will be able to receive only a portion of the social returns to intangible capital. As a result, the entrepreneur chooses a lower level of intangible investment and/or a smaller span of production streams. For society as a whole, however, it would be preferable to have a higher level of investment and its positive spillovers.

Stage 1: Intangible Investment/Creation

The first step in the timeline of our model is to determine the initial investment in the intangible asset. Here, we can think of the entrepreneur as exerting effort in search of a profitable new idea, which will be for the new intangible. Higher effort is more likely to yield more profitable ideas, but a substantial level of risk is involved. For present purposes, we are especially interested in how the entrepreneur's choice of effort depends on the parameters governing limited excludability and the degree of non-rivalry in use.

If excludability is low, then an entrepreneur will have less incentive to exert effort in generating intangibles, because their ideas can be expropriated. By contrast, the effects of the non-rivalry parameter on effort are more subtle, in a way related to the complementarity we have emphasized between non-rivalry and excludability. The scalability of creating intangible assets may generate value, but the entrepreneur will only value the associated intangible asset to the extent that the benefits are sufficiently excludable. Also as noted earlier, the model features underinvestment in intangibles since the entrepreneur's effort choice depends on the *private* value of the intangible, which in general is lower than the social value, or the value to outside investors. Perhaps surprisingly, the degree of under-investment can be greater for intangibles that are highly scalable if excludability is low enough. The intuition follows from the

complementarity argument above: since highly scalable intangibles only generate value to the entrepreneur if they are appropriable, the entrepreneur will especially undervalue highly scalable intangibles when they lack appropriability.

Finally, it is important to emphasize the distinction between expected and realized returns to the entrepreneur. If there is selection on which entrepreneurs enter the market (or equivalently if failure in creating an intangible asset is a feasible outcome despite the amount of effort involved) then focusing on compensation received by entrepreneurs will be misleading. With free entry of entrepreneurs, it is possible to have a situation in which the expected rents from creating intangible capital are zero, but the realized rents from doing so—looking only at those who succeeded—can be positive. This is often emphasized in analysis of patents, for example, where the observed payoffs may be high but may not fully capture the risk and the failures that are known before the patent is created but may be unobserved by looking at patent success stories.

Implications and Relation to Other Approaches

Our approach to intangibles, incorporating non-rivalry and limited appropriability, leads to unique implications compared to the standard neoclassical model. We now discuss the relationship between our model and the key properties of intangibles used in existing work, especially with regard to non-rivalry and limits to excludability.

A common premise in the literature on intangible capital is that it can contribute to "higher returns to scale," or more generally, that intangibles are "scalable" (Haskel and Westlake 2018). A standard rationale is that to assume that intangible investment involves high fixed costs, but leads to lower marginal cost of production for the firm. As a result, production at intangible-intensive firms may be characterized by increasing returns to scale, at least locally. Of course, locally increasing returns are not specific to intangibles; for instance, the production of power from nuclear plants relies heavily on physical inputs but has the same profile of high fixed and low marginal costs.

Instead of assuming a particular structure of fixed and marginal costs that lead to increasing returns, our model starts from the idea of non-rivalry in use. As a result, our notion of "scalability" is somewhat different from the existing literature. In our model, "scalability" derives endogenously from the complementarity between intangible capital and firm span, which arises so long there is some degree of non-rivalry in use within the firm. In this case, the value of the intangible asset increases when it can be employed in multiple segments. Similarly, the higher the value of the firm's intangible asset, the greater is the benefit of expanding the span of operations.

This scalability property is modulated by limits to the excludability of returns from investment in intangible capital. Here, we emphasize that greater excludability is not always desirable. The flipside of limits to excludability is that intangibles can generate spillovers outside of the firm. Ideas that are stored and widely disseminated can be used effectively in production—or even spur the development of better ideas. The use of a specific intangible asset by one firm may indirectly increase

productivity in other firms who can potentially adopt the same intangible. Negative spillovers are also possible. The same forces that lead to wide dissemination and adoption of new ideas imply that older ideas become more easily obsolete. A new and more efficient method of production can be licensed to many firms, leading to a drop in the value of the intangible asset (say, a patent) representing the old production method. The assumption that there are limits to excludability could also capture this process of "external" depreciation resulting from a firm's investment in intangible assets (Jovanovic and Rousseau 2002).

Intangible Capital and Economic Trends

Economists typically estimate output and the stock of physical capital with greater precision than the stock of intangible assets. Indeed, measuring the stock of intangibles as an input to production is quite challenging, since they can be embodied in a variety of media, including human capital. However, accounting for intangible assets and understanding their unique characteristics can shed light on some key economic trends. In particular, the period since the 1990s has been characterized by relatively low growth in total factor productivity, a decline in the labor share, weak tangible investment and rising valuations, an increase in economic rents, and rising inequality.

The Productivity Slowdown

After a productivity boost in the 1990s, the United States (and other) economies have seen a widespread productivity slowdown during the last two decades. Based on a standard aggregate production function, growth in output can be decomposed into growth of each of the inputs, like capital and (quality-adjusted) labor. The unexplained "Solow residual" term is then taken as a measure of the change in total factor productivity. But if a substantial part of intangible capital is not captured in the statistics on "capital" inputs—and indeed, some of intangible capital investment is embodied in wages paid to, say, those creating intangible capital—can this help to explain the decline in measured productivity?

One can take two approaches in re-interpreting the official measures of productivity with intangible capital in mind: reinterpreting the Solow residual and re-estimating intangible capital.

The first approach would treat the *entirety* of the Solow residual as driven by incomplete measurement of intangible capital. In this view, intangibles and their properties are the "dark matter" that explain a wedge in between measured output and inputs. This is essentially the view adopted by some models of endogenous growth (for a recent survey, see Jones 2021). In this view, slower productivity growth could result from lower investment in intangibles. It could also result from changes in the degree of non-rivalry or appropriability, if the spillovers from intangible capital to the rest of the economy decreased. From this approach, the slowdown in the growth rate of measured productivity can be rationalized in three possible ways.

First, the benefits of intangible capital may be delayed by substantial *time-to-build*: as noted earlier, its effects may resemble types of tangible capital with high fixed costs of installation, but with negligible variable costs. For example, logistics-optimization software may require substantial development time, but once operational, it can improve delivery times for all of a firm's production units. New production methods may take time to be adopted and for learning-by-doing to take effect. To the extent that these up-front costs are not recognized as investment, they can generate a slowdown in measured productivity over the short- and even the medium-run even as long-run productivity is higher (Brynjolfsson, Rock, and Syverson 2021).

Second, obsolescence of intangible capital may obscure measured total factor productivity growth. Because intangible capital can be superseded by innovation, it can become obsolete quickly. Combined long time-to-build lags and displacement of existing intangible capital during periods of rapid innovation can exacerbate the slow rise of measured output in the short run (Greenwood and Jovanovic 1999).

Third, the fact that the investing firm or entrepreneur cannot capture all of the value of intangible investment may reduce the incentive to create intangible capital internally. As we discussed earlier, the degree of non-rivalry or degree of appropriability not only affect output directly, but they also indirectly affect the incentive to exert effort to create new intangibles. Variations in the effective degree of appropriability can also lead to fluctuations in measured total factor productivity (as in the model of Kondo, Li, and Papanikolaou 2021).

The second overall approach to re-interpreting the official measures of productivity with intangible capital in mind re-estimates the stock of intangibles, rather than reinterpreting the existing Solow residual. In this approach, researchers aim to construct more "complete" measures of capital, in part by estimating intangible capital directly, and then to see how the total factor productivity residual adjusts. This approach proceeds by identifying expenditures and prices of excluded intangible investment, and then using them to construct estimates of the intangibles stock, sometimes with the help of an equilibrium model. For example, this approach is followed in Basu, Fernald, Oulton, and Srinivasan (2003) and Corrado, Hulten, and Sichel (2005), and also explored in McGrattan and Prescott (2010a, 2010b), and Crouzet and Eberly (2021a).

One lesson from this literature is that including unmeasured intangibles has opposing effects on the Solow residual measure of total factor productivity growth. For example, McGrattan and Prescott (2010b) argue that intangible capital is typically expensed (that is, treated as a production cost rather than as an investment), or else is financed by employees' "sweat equity." They show how the resulting underestimation of output and income, due to not measuring the production of investment in the form of intangible capital, mechanically leads to lower measured labor productivity growth in the 1990s. The effect on total factor productivity growth is less clear: underestimated intangible capital becomes both an additional output (which raises actual total factor productivity) and an additional input to production (lowering actual total factor productivity). These two opposing effects imply that

the exact timing of the additional input and the additional output affects the path of estimated total factor productivity.

This literature, as well as national statistical agencies, also faces open questions on how to measure intangible capital: for example, how to construct appropriate price indices to deflate past investment expenditures in intangibles. A further complication is that output, intangibles, and measured productivity in the United States may be mismeasured for other reasons; for example, due to the fact that US corporations have a tax incentive to book income from intangible assets abroad (Guvenen et al. 2021). The income that is booked offshore lowers gross operating surplus in the United States, thereby reducing measured value added. The impact on measured total factor productivity depends on the countervailing effects of missing intangible inputs and missing income. This is less of an issue in accounts that consolidate firms' activities across countries, but it is a question for national-boundary-based measures.

From the perspective of our model, a main difference with this literature is that it treats intangible investment as an input with similar economic characteristics to physical capital; in effect, it assumes that intangible and physical capital are substitutes along with other factors of production. However, a fundamental property of intangibles is non-rivalry, which together with limited appropriability, can lead to positive spillovers and thus also affect measured productivity. To measure the full contribution of intangibles to economic output, researchers need to measure not only the intangible stock but also to account for the value of spillovers, which involves estimation of the parameters governing non-rivalry in use and limited appropriability.[1]

Factor Shares

The labor share of income in national accounting data has been declining, both in the United States and globally (for example, Elsby, Hobijn, and Şahin 2013; Karabarbounis, Loukas, and Neiman 2014). When interpreting these trends, however, the existence of intangibles implies that factor shares are also mismeasured. Depending on the implicit assumptions researchers make, the share of output that would accrue to intangible inputs could be allocated to either physical capital, labor, or "rents," where the latter is defined as monopoly profits.

Our model helps shed some light on the underlying issues. Recall that the limits to excludability can be motivated by incomplete markets, in which the entrepreneur gives up some rents by finding a way to store some of the intangible externally, in exchange for funds for expansion from outside investors. In addition, at an optimum, the greater the degree of non-rivalry, the smaller the share that accrues to the entrepreneur. The entrepreneur chooses to give up a larger fraction of rents

[1] To the extent that aggregate market values are used to measure the price of intangibles included in national accounts, some of these spillovers may be included in existing estimates of the intangible capital stock. But aggregate market values for intangibles are scarce, and even when these measures exist, it is not always clear how to allocate spillover returns across sectors (Moylan and Okubo 2020). Further, market values may underestimate the contribution of intangibles if these are partly stored in key labor inputs (Eisfeldt and Papanikolaou 2014).

to achieve a higher span of product streams (and scale). Should the entrepreneur's remaining share be treated as returns to capital or to labor? If human capital is the key input in the production of new intangibles, it is labor income and hence part of labor share. The residual to which outside investors have a claim could be (though it need not be) part of capital income. Given that many intangible assets like patents, copyrights, and trademarks confer exclusivity and hence monopoly power, the conceptual distinction between monopoly rents and the factor share of intangibles may be hard to disentangle.

Cash flows from intangibles likely go to both labor (key talent) as well as owners of tangible capital (shareholders). The appropriate allocation likely varies across different types of intangibles, depending on how the intangible capital is stored, and then on its property rights and excludability. For example, managers may accrue income from intangible assets such as organization processes or corporate culture. Similarly, capital owners (shareholders) may accrue income from software, patents, or brands. In general, the full value of intangible capital will be observed partly on the market value balance sheet of firms, and partly on the market value balance sheet (wealth) of key talent. Variation in bargaining power between the two parties can lead to insiders appropriating a larger or smaller share. As a result, imputing the stock of intangibles based on firms' market valuation ratios will likely underestimate the value of intangibles (Eisfeldt and Papanikolaou 2014).

Any calculation of labor and capital factor shares should make these distinctions, but many do not. For instance, Barkai (2020) measures monopoly rents as output minus labor expenses minus the stock of physical capital times its user cost. As a result, this measure of monopoly rents will include the share of output due to intangibles. Alternatively, if one estimates the capital share as one minus the wage payments to labor (Elsby, Hobijn, and Şahin 2013; Karabarbounis and Neiman 2014), this factor share for capital will include income from intangibles. If labor is a key input in the production of intangibles, part of that income should in fact be classified as labor income. To the extent that intangibles are stored in key employees, this choice can understate the labor share. In particular, some capital income and profits are actually equity compensation for high-skilled labor inputs and are thus partially misclassified. Eisfeldt, Falato, and Xiaolan (2021) document the large fraction of labor compensation in the form of equity-based pay in recent decades. Ownership of private firms can lead to compensation of labor inputs with capital income (Smith et al. 2019; Bhandari and McGrattan 2021). Note also that factor shares have an important impact on measuring total factor productivity, which we discussed in the previous section. In addition to the direct effect of measuring factor growth, misclassification of intangibles as intermediate inputs can bias factor shares and reduce estimates of total factor productivity (Crouzet and Eberly 2021a).

A further complication in interpreting trends in the labor share is that total output is also mismeasured if expenditures on intangibles are recorded as a cost of production, rather than investment, as we noted in our discussion of productivity measurement. Indeed, expenditures on intangibles tend to be recorded either as payments to labor or as purchases of intermediate inputs, such as consulting

services or lab equipment.[2] Recent revisions to the US national income and product accounts capitalized certain intangible assets, especially intellectual property products, and added these "produced" assets to gross operating surplus and as income to capital. Indeed, this allocation of intellectual property products to the capital account in the US national income and product accounts generated most of the measured decline in the labor share (Koh, Santaeulàlia-Llopis, and Zheng 2020).

Because current practice tends to either omit intangibles from national accounts or to allocate their payments primarily to capital, the actual labor share is likely higher than the share computed using national income and product data (in particular after the statistical revisions in 1999 and 2013 that treated all of intellectual property as capital).

Inequality of Income and Wealth

Just as the treatment of intangibles can bias our view of capital and labor shares, the rising importance of intangible assets can lead to inequality between those who benefit from intangibles and those who do not. Here we discuss three mechanisms by which this could occur: i) rents may accrue to inventors and entrepreneurs, ii) rents may accrue to key employees, and iii) intangibles may exacerbate capital-skill complementarity.

Under the first mechanism, when inventors or entrepreneurs conceive and develop a new intangible, they can appropriate a fraction of the value generated. The rest of the value generated accrues to outsiders, including outside investors, or other firms and consumers in the economy more broadly. The key difference between these two parts is that the entrepreneur's share is concentrated and not easily tradeable. Concentrated exposure to intangibles can lead to inequality through both the drift in the owner's wealth (if the intangible is exposed to substantial systematic risk) and from the idiosyncratic shocks to the intangible's value. The entrepreneur cannot pre-sell claims to future intangibles that have yet to be produced (otherwise, the incentive to exert effort in creating such intangibles after already receiving payment would be low). These early-stage intangibles will have very concentrated ownership. By contrast, by codifying and storing the intangible, the entrepreneur creates an asset to which outsiders can lay a claim. Moreover, these claims can be by diffuse investors who can build diversified portfolios.

This key distinction between the two shares—what is stored with the entrepreneur versus what is owned by outside stakeholders—can help to explain rising inequality, as Kogan, Papanikolaou, and Stoffman (2020) explore in a general equilibrium model. The key feature of their model is incomplete markets: during each period, a small measure of agents—the "inventors"—are randomly endowed with a

[2]A substantial fraction of investment in research and investment is labor compensation. For example, Adobe's research and development expenditures in 2019, and Sanderson Farm's breakdown of Selling and General Administrative Expenses in 2019, as shown by the respective 10-Ks of these firms, show that the majority of their research and Selling and General Administrative expenditures are in fact labor compensation.

blueprint for a new project. They interpret these inventors broadly as encapsulating all parties that share the rents from new investment opportunities, other than the owners of the firm's publicly-traded securities. This model suggests that firm owners as a group reap only part of the benefits of innovation, but bear all the costs of creative destruction. As a result, the arrival and churn in new technologies is associated with greater income and wealth inequality, together with a motive to insure against states of the world with rapid technological innovation.

The second mechanism begins with the insight that while part of investment in an intangible asset is codified in media, another part could be stored with key employees. An example would be management practices: a textbook or a business school can prescribe "best practices," but it does not immediately follow that everyone who takes the class or reads the book becomes an effective manager. Further, because the value of the intangibles stored with key employees are likely to be partly specific to the firm, the cash flows generated by this form of intangible capital is often shared between shareholders and key talent. As intangibles grow in importance in terms of firms' capital stocks and value, the importance of key labor inputs in contributing to and operating these assets may lead to inequality between key labor and other workers (McGrattan and Prescott 2010b; Smith et al. 2019; Eisfeldt, Falato, and Xiaolan 2021; Bhandari and McGrattan 2021). Further, the share of rents that accrues to key talent need not be constant, as it depends on their outside option, which introduces a further source of inequality (Eisfeldt and Papanikolaou 2013).

The third mechanism we consider is how intangibles can affect income inequality even when intangibles are all stored externally, if they exacerbate capital-skill complementary. A common view of technological progress since the mid-twentieth century is that it is primarily skill-biased—that is, technology is generally a complement to high-skill workers, but was a substitute for low-skill workers (Goldin and Katz 2008). If intangibles increase the marginal productivity of high-skilled labor inputs, then the rise in the importance of intangibles over the last few decades may have contributed to rising inequality. In support of this view, Eisfeldt, Falato, and Xiaolan (2021) find that pay to high-skilled labor, and in particular equity pay, grew fastest in recent decades within industries which were most exposed to declining investment goods prices, as a proxy for the growth in intangibles stored in capital goods. Although the model we developed earlier in this paper has no explicit labor inputs, it is relatively straightforward to include different types of labor into the model, possibly using a constant-elasticity-of-substitution framework similar to Krusell et al. (2000) and Eisfeldt, Falato, and Xiaolan (2021).

While the above three mechanisms imply that rising intangibles lead to higher inequality, the relation between intangibles and the level of income inequality can be ambiguous if new intangibles are also associated with skill displacement. Put differently, an increase in between-group inequality need not translate into an increase in between-worker inequality if workers transition across groups. Using detailed data on patent inventions and occupation task descriptions, Kogan et al. (2021) document that workers in occupations most exposed to technology improvements tend to experience declines in wage earnings. They find that the workers most adversely

affected are the highest-paid workers, which can be consistent with technology-skill complementarity only as long as workers can also lose part of their human capital when technology improves.

Tangible Investment and Tobin's *q*

In recent years, investment in physical capital has been declining, while measures of the return to physical capital have been rising. Figure 3 illustrates this trend with data from US public non-financial firms. The blue line is the aggregate investment rate in physical capital. The orange line measures the rate of return to investment, measured as Tobin's *q*—the ratio of total enterprise value (the sum of the value of equity plus debt, adjusted for liquid asset holdings), to the stock of physical assets. The two lines show a positive correlation at business-cycle frequencies (as highlighted in Andrei, Mann, and Moyen 2019). However, over the longer run, they diverge: the investment rate fell by about 4 percentage points, while Tobin's *q* increased by a factor of about 3.

For models of investment where firms only rely on physical capital (Hayashi 1982; Abel and Eberly 1996), explaining this long-run divergence is a challenge, as first noted by Gutiérrez and Philippon (2016). According to these models, Tobin's *q* should proxy for the marginal benefits of physical investment, because firm value scales proportionally to the stock of physical capital.[3] Higher Tobin's *q* should signal higher returns to physical capital and encourage physical investment, at odds with the trend in Figure 3.

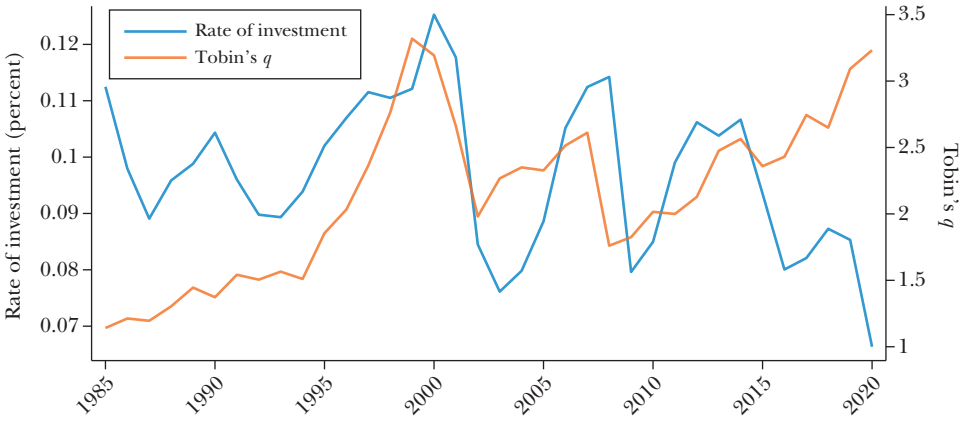
Intangible capital, using the model developed earlier, can shed light on this puzzle in at least two ways.[4] First, with intangibles, the numerator of Tobin's *q* captures the market value of the whole firm, including benefits of intangible capital and any spillovers. The denominator, on the other hand, is only physical capital. Hence, the ratio overstates the true return to physical capital to the extent that value is also generated by intangibles. Put differently, with intangibles, average returns to physical capital will overstate the true incentive to invest in physical capital.[5] The bias will grow as firms increase intangible inputs relative to physical capital, as the evidence in the introduction suggests they have in recent decades. Consistent with this intuition, Crouzet and Eberly (2019) show that, controlling for the ratio of intangible to physical capital, the difference in trends between measured returns and investment rates for physical capital highlighted in Figure 3 shrinks by about 30 percent.

[3]Tobin's *q* is an empirical measure of the average return to physical capital: total enterprise value per unit of physical capital. Moreover, in traditional models, the average and the marginal return to physical capital should be the same, because the firm is assumed to be able to scale up revenues one-for-one with physical capital. Finally, the marginal return to physical capital—how much enterprise value rises for each incremental unit of physical capital—measures the marginal benefit of investing in physical assets. As a result, investment in physical capital should be tightly linked to Tobin's *q*.

[4]The online Appendix to the paper, as well as the companion paper (Crouzet et al. 2022b), provide a more formal discussion of these two points.

[5]This argument was first outlined by Hayashi and Inoue (1991).

Figure 3
Physical Capital: Rate of Investment and Return on Investment



Source: Author's calculations. See online Appendix for details.
Note: The data source is the sample of US non-financial corporations in Compustat. The gross physical investment rate is the ratio of aggregate capital expenditures (Compustat item capx) to aggregate, gross property, plant, and equipment (Compustat item ppegt). The return on investment reported is Tobin's *q*: the ratio of enterprise value—the aggregate market value of equity (the product of Compustat items prc and shrout), plus the aggregate book value of debt (the sum of Compustat items dlc and dltt), minus cash and cash equivalent (Compustat item che), to aggregate gross property, plant, and equipment.

Second, for a given ratio of intangible to physical capital, changes in the degree of non-rivalry and excludability of intangibles will also affect the relationship between measured average returns and investment rates for physical capital. With a higher degree of non-rivalry within the firm, marginal returns to intangible investment will rise, so that more of total enterprise value will be accounted for by intangibles. An example could be the growing availability (and declining price of) of digital media, which facilitate the replication and scaling of intangibles stored in software. Similarly, higher excludability (say, from strengthening enforcement of property rights institutions, like patents) would increase the share of cash flows that can be retained by firm owners, and therefore the contribution of intangibles to total enterprise value. In both cases, the gap between measured returns to physical capital and the true marginal product will rise (or conversely would fall with weaker non-rivalry and excludability). These forces could also explain the trend presented in Figure 3.

The wedge between average and marginal returns to physical capital can be affected by a rising share of intangibles in other ways, as well. As we noted in our discussion of productivity measurement, the arrival of new intangible assets (say new designs) may reduce the value of existing physical capital, so that average returns to investment could fall even as marginal returns to investment rise. As an illustration, Greenwood and Jovanovic (1999) found that the arrival of information technology led to a fall in the value of existing firms. In a model with technology

shocks embodied in new capital goods (for example, Papanikolaou 2011), improvements in new capital goods can lead to a decline in the average rate of return if the replacement value of the installed capital stock (the denominator in calculating Tobin's *q*) does not adjust fully to offset the decline in the market value of incumbent firms. In this case, one could observe rising marginal returns and new investment together with weak Tobin's *q*.

Rents and Market Structure

We define rents as returns generated by an asset in excess of its marginal (user) cost. Like physical capital, intangible capital can generate rents for its owners. Rents relate to market structure because they often arise in situations where capital is used to produce a good for which there are poor substitutes. For instance, a drug formula produces rents to the extent that it is difficult to produce a generic substitute. These conditions can lead to imperfect competition and, in some cases, to greater market concentration.

In situations where intangible capital generates rents, the rent is the stream of appropriable returns generated by the intangible in excess of its user cost. In a business franchise, for example, the intangible asset is the combination of the brand with the logistical and organizational instructions provided to the franchisee. To generate rents, the asset must produce returns that exceed cost of using it (paying the franchiser, implementing organizational instructions, and possibly further promoting the brand). It may be possible for the franchiser and franchisee to appropriate these rents—for instance, through enforceable franchise agreements.

Conceptually, the rents generated by an intangible or tangible asset should be measured separately from the intangible itself. In this approach, the value of a firm can be divided into four categories: the value of physical capital, value of rents from physical capital, value of intangible capital, and value of rents from intangible capital. Crouzet and Eberly (2021b) show that this decomposition holds in a broad class of dynamic investment models in which capital inputs, both physical and intangible, can generate rents. Moreover, they describe how to estimate the components of the decomposition using a set of statistical moments, including Tobin's (average) *q* for physical assets, flow returns to physical capital, and an estimate of the ratio of intangible to physical capital. The key finding is that rents associated with intangible assets have contributed to a sharply rising share in the growth of total enterprise value of US businesses since the early 1990s, accounting for approximately 15 percent in the mid-1980s and to up to 40 percent in 2015, depending on how broadly intangibles are measured.[6]

[6]Crouzet and Eberly (2021b) allow for intangibles and market power within a neoclassical framework which does not incorporate non-rivalry of intangibles nor for limits to excludability. Allowing for the features we introduce here, non-rivalry of intangibles would tend to increase rents, while limits to excludability would reduce them. Thus, one should expect industries in which intangibles are highly non-rival but easy to exclude to have particularly high rents from intangibles. A potential example is the health care industry, where intangible assets (like drug patents) are easy to replicate within the firm but well

As discussed earlier, when creating intangibles involves fixed costs, their non-rivalry within the firm can generate scale economies. Scale economies in turn may lead to higher market concentration. Note that this does not necessarily require that firms earn rents in the first place, so that higher concentration need not go hand in hand with more rents. Instead, the effects of non-rivalry on concentration are closely related to the theory of natural monopolies, which also emphasizes that markets featuring some firms with high-fixed, low-marginal cost structures will have high equilibrium concentration (Baumol 1977). Thus, more intangible-intensive industries may naturally be more concentrated, more intangible-intensive firms should command higher market shares, and intangible intensity within firms should be correlated with market share. Crouzet and Eberly (2019) provide reduced-form evidence consistent with these predictions, using data on publicly traded US firms.[7]

Intangibles might also be conducive to higher concentration by encouraging consolidations across firms. Non-rivalry implies that, rather than having two firms each bear the fixed cost of creating the same intangible asset, it may be efficient for them to merge and share the cost. For instance, two retailers might merge and operate under the same brand, rather than creating and promoting their brands separately. More broadly, the non-rival nature of intangibles may play into determining the boundaries of the firm.

Implications for Corporate Finance and the Cost of Capital

Even when intangible assets can be stored, limited excludability implies that intangibles are less likely to be pledgeable to outside investors than physical capital. Lack of pledgeability undermines the viability of debt contracts, so that debt is less likely to be the preferred form of financing (Falato et al. 2020). An alternative possibility, Sun and Xiaolan (2019) argue, is that intangibles are primarily financed by employees, who then have an implicit claim on them. Related to this point, the capital structure of new firms with substantial intangibles appears to include more concentrated control in the form of dual-class shares whereby initial owners, founders, and employees retain more voting rights than outside equity holders (Ritter 2022).

The presence of intangibles can also affect the cost of capital. Given that intangibles may be stored in non-capital inputs, the sharing rule for cash flows generated by intangible capital may also expose investors in firms relying on intangible assets to unique risk factors (Eisfeldt and Papanikolaou 2013; Eisfeldt, Kim, and Papanikolaou 2020).[8] Further, if the economic value that is generated by new ideas cannot be fully pledged and hence diversified to outside investors, then states of the world with rapid technological innovation will be associated with higher inequality, creating a

protected by property rights institutions. The online Appendix includes an algebraic derivation of these results.

[7] Relatedly, Kwon, Ma, and Zimmermann (2022) provide long-run evidence on the relationship between the degree of industry concentration and expenditures on intangibles such as research and development capital.

[8] See also Hansen, Heaton, and Li (2005) for the effect of intangibles on valuation.

demand for insurance against these states and leading to a lower cost of capital for fast-growing firms (Kogan, Papanikolaou, and Stoffman 2020).

Concluding Thoughts

We have sought to describe intangibles affirmatively—rather than simply a lack of physical form—as a way of illustrating their role in production more clearly. This topic offers many opportunities for future research. Understanding how measured productivity can remain so weak in the face of seemingly continuous innovation remains puzzling. Further research on measurement and the interaction between intangibles and other factors, especially through non-rivalry, may shed light on this apparent contradiction. The factor income earned by intangibles is particularly difficult to classify as capital or labor income, or to allocate within types of labor. Further refinement of the allocation of rents accruing to intangibles may illuminate sources of market power, especially arising via the excludability conferred on intangibles through patents, copyrights, and other institutions. The distribution of these rights may also shed light on the sources of rising inequality. Further exploration of potential connections between intangibles, market power, and industry concentration is a rich area for future research.

Moreover, we suspect that the rise of intangible assets may spur institutional developments. As we noted, firms are relying more on intangible capital for production, yet many types of intangible capital do not yet have an institutional framework to use it as collateral for financing. Some intangibles have lively secondary markets: as one example, in licensing of patents and copyrights. Other intangibles seem less separable from the rest of firm, and the institutional framework is less clear: for example, some states have widespread non-compete agreements, based on a belief that such agreements provide greater excludability for certain kinds of intangible capital, while others do not. Economists may have much to contribute to developing and implementing the potential tradeoffs between different institutional frameworks that affect incentives to invest in intangible capital.

■ *We thank Andrew Atkeson and Mindy Xiaolan for helpful discussions and the editors for their guidance and helpful comments. Also, we thank Edward Kim (UCLA Anderson PhD student) for helpful research assistance.*

References

Abel, A.B., and Eberly, J.C. 1996. "Optimal Investment with Costly Reversibility." *Review of Economic Studies* 63 (4): 581–93.

Aggarwal, Dhruv, Ofer Eldar, Yael V. Hochberg, Lubomir P. Litov. 2022. "The Rise of Dual-Class Stock IPOs." *Journal of Financial Economics* 144 (1): 122–53.

Andrei, Daniel, William Mann, and Nathalie Moyen. 2019. "Why Did the q Theory of Investment Start Working?" *Journal of Financial Economics* 133 (2): 251–72.

Barkai, Simcha. 2020. "Declining Labor and Capital Shares." *Journal of Finance* 75 (5): 2421–63.

Basu, Susanto, John G. Fernald, Nicholas Oulton, and Sylaja Srinivasan. 2003. "The Case of the Missing Productivity Growth." *NBER Macroeconomics Annual* 18: 9–63.

Baumol, William J. 1977. "On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry." *American Economic Review* 67 (5): 809–22.

Benmelech, Efraim. 2009. "Asset Salability and Debt Maturity: Evidence from Nineteenth-Century American Railroads." *Review of Financial Studies* 22 (4): 1545–84.

Bhandari, Anmol, and Ellen R. McGrattan. 2021. "Sweat Equity in US Private Business." *Quarterly Journal of Economics* 136 (2): 727–81.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2021. "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies." *American Economic Journal: Macroeconomics* 13 (1): 333–72.

Corrado, Carol, Charles Hulten, and Daniel Sichel. 2005. "Measuring Capital and Technology: An Expanded Framework." In *Measuring Capital in the New Economy*, edited by Carol Corrado, John Haltiwanger, and Dan Sichel, 11–46. Chicago: University of Chicago Press.

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

Crouzet, Nicolas, and Janice Eberly. 2019. "Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles." Paper presented at the Economic Policy Federal Reserve Bank of Kansas City Symposium, Jackson Hole, WY, August 23, 2018.

Crouzet, Nicolas, and Janice Eberly. 2021a. "Intangibles, Markups, and the Measurement of Productivity Growth." *Journal of Monetary Economics* 124 (S): 92–109.

Crouzet, Nicolas, and Janice Eberly. 2021b. "Rents and Intangible Capital: A Q+ Framework." NBER Working Paper 28988.

Crouzet, Nicolas, Janice Eberly, Andrea Eisfeldt, and Dimitris Papanikolaou. 2022a. "Replication data for: Econonomics of Intangible Capital." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E172021V1.

Crouzet, Nicolas, Janice Eberly, Andrea Eisfeldt, and Dimitris Papanikolaou. 2022b. "A Model of Intangible Capital." Unpublished.

Eisfeldt, Andrea L., and Dimitris Papanikolaou. 2013. "Organization Capital and the Cross-Section of Expected Returns." *Journal of Finance* 68 (4): 1365–406.

Eisfeldt, Andrea L., and Dimitris Papanikolaou. 2014. "The Value and Ownership of Intangible Capital." *American Economic Review* 104 (5): 189–94.

Eisfeldt, Andrea L., Edward T. Kim, and Dimitris Papanikolaou. 2022. "Intangible Value." *Critical Finance Review* 11 (2): 299–332.

Eisfeldt, Andrea L., Antonio Falato, and Mindy Z. Xiaolan. 2021. "Human Capitalists." NBER Working Paper 28815.

Elsby, Michael W.L., Bart Hobijn, and Aysegül Şahin. 2013. "The Decline of the U.S. Labor Share." *Brookings Papers on Economic Activity* (2): 1–63.

Falato, Antonio, Dalida Kadyrzhanova, Jae Sim, and Roberto Steri. 2020. "Rising Intangible Capital, Shrinking Debt Capacity, and the US Corporate Savings Glut." Unpublished.

Goldin, C.D., and L.F. Katz. 2008. *The Race Between Education and Technology*. Cambridge, MA: Harvard University Press.

Greenwood, Jeremy, and Boyan Jovanovic. 1999. "The Information-Technology Revolution and the Stock Market." *American Economic Review* 89 (2): 116–22.

Gutiérrez, Germán, and Thomas Philippon. 2016. "Investment-less Growth: An Empirical Investigation." Unpublished.

Guvenen, Fatih, Raymond J. Mataloni, Jr., Dylan G. Rassier, and Kim J. Ruhl. 2021. "Offshore Profit Shifting and Aggregate Measurement: Balance of Payments, Foreign Investment, Productivity, and the Labor Share." Unpublished.

Hall, Robert E. 2001. "The Stock Market and Capital Accumulation." *American Economic Review* 91(5): 1185–202.

Hansen, Lars Peter, and John C. Heaton, and Nan Li. 2005. "Intangible Risk." In *Measuring Capital in the New Economy*, edited by Carol Corrado, John Haltiwanger, and Daniel Sichel, 111–52. Chicago: University of Chicago Press.

Haskel, Jonathan, and Stian Westlake. 2018. *Capitalism without Capital: The Rise of the Intangible Economy*. Princeton: Princeton University Press.

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

Hayashi, Fumio, and Tohru Inoue. 1991. "The Relation between Firm Growth and Q with Multiple Capital Goods: Theory and Evidence from Panel Data on Japanese Firms." *Econometrica* 59 (3): 731–53.

Jones, Charles I. 2021. "The Past and Future of Economic Growth: A Semi-Endogenous Perspective." Unpublished.

Jovanovic, Boyan, and Peter L. Rousseau. 2002. "The Q-Theory of Mergers." *American Economic Review* 92 (2): 198–204.

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

Kermani, Amir, and Yueran Ma. 2020. "Asset Specificity of Non-Financial Firms." NBER Working Paper 27642.

Kogan, Leonid, Dimitris Papanikolaou, and Noah Stoffman. 2020. "Left Behind: Creative Destruction, Inequality, and the Stock Market." *Journal of Political Economy* 128 (3): 855–906.

Kogan, Leonid, Dimitris Papanikolaou, Lawrence D.W. Schmidt, and Bryan Seegmiller. 2021. "Technology-Skill Complementarity and Labor Displacement: Evidence from Linking Two Centuries of Patents with Occupations." NBER Working Paper 29552.

Koh, Dongya, Raül Santaeulália-Llopis, and Yu Zheng. 2020. "Labor Share Decline and Intellectual Property Products Capital." *Econometrica* 88 (6): 2609–28.

Kondo, Jiro, Danielle Li, and Dimitris Papanikolaou. 2021. "Trust, Collaboration, and Economic Growth." *Management Science* 67 (3): 1825–50.

Krusell, Per, Lee E Ohanian, José-Víctor Ríos-Rull, and Giovanni L. Violante. 2000. "Capital-Skill Complementarity and Inequality: A Macroeconomic Analysis." *Econometrica* 68 (5): 1029–53.

Kwon, Spencer Yongwook, Yueran Ma, and Kaspar Zimmermann. 2022. "100 Years of Rising Corporate Concentration." Unpublished.

McGrattan, Ellen R., and Edward C. Prescott. 2010a. "Technology Capital and the US Current Account." *American Economic Review* 100 (4): 1493–522.

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

Moylan, Carol E., and Sumiye Okubo. 2020. "The Evolving Treatment of R&D in the U.S. National Economic Accounts." *BEA Papers*, March 2020, 1–29.

Papanikolaou, Dimitris. 2011. "Investment Shocks and Asset Prices." *Journal of Political Economy* 119 (4): 639–85.

Ramey, Valerie A., and Matthew D. Shapiro. 2001. "Displaced Capital: A Study of Aerospace Plant Closings." *Journal of Political Economy* 109 (5): 958–92.

Ritter, Jay R. 2022. "Initial Public offerings: Dual Class Structure of IPOs through 2020." Unpublished.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2019. "Capitalists in the Twenty-First Century." *Quarterly Journal of Economics* 134 (4): 1675–745.

Sun, Qi, and Mindy Z. Xiaolan. 2019. "Financing Intangible Capital." *Journal of Financial Economics* 133 (3): 564–88.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 53–74

Marketing Investment and Intangible Brand Capital

Bart J. Bronnenberg, Jean-Pierre Dubé, and Chad Syverson

During the early and mid-twentieth century, several economists called for the systematic study of the social benefits of the rapid escalation in marketing: advertising, branding, promoting, selling, and trademarking (for example, Shaw 1912). Braithwaite (1928, p. 16) observed that "goods cost as much to market as they do to manufacture." Coase (1937, p. 394) went so far as to conclude "the introduction of the firm was primarily due to the existence of marketing costs." According to the *persuasive* or *prestige* view of marketing efforts—around since at least Marshall (1919) and argued famously by Bain (1956) and Galbraith (1958)—brand-driven preferences might make consumers less price-elastic, creating barriers to entry and raising the market power of incumbent firms. Moreover, if advertising by one firm does nothing but offset advertising by other firms, then such spending may be socially wasteful.

However, an alternative view of marketing gradually emerged. Under the *informative* view, as in Stigler (1961) or Telser (1964), marketing serves primarily to make consumers aware of the product and/or of its attributes. Such information reduces

■ *Bart J. Bronnenberg is Professor of Marketing, Tilburg University, Tilburg, Netherlands. Jean-Pierre Dubé is James M. Kilts Distinguished Service Professor of Marketing and Chad Syverson is George C. Tiao Distinguished Service Professor of Economics, both at the University of Chicago Booth School of Business, Chicago, Illinois. Bronnenberg is also a Research Fellow, Centre for Economic Policy Research, London, United Kingdom. Dubé and Syverson are both Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are Bart.Bronnenberg@tilburguniversity.edu, Jean-Pierre.Dube@chicagobooth.edu, and 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.36.3.53.

search costs and makes markets more competitive. Under a related *complementary* view, consumption of known brands enters directly into the utility function. These approaches suggest that marketing can raise welfare.

Although the dispute between these views has neither been resolved nor completely died out, we believe the more recent economics literature has unduly neglected intangible marketing and brand capital and its many micro and macro implications in its studies of industry structure, productivity, and aggregate output. For mostly technical reasons, the emerging literature on general equilibrium analysis during the mid-twentieth century worked with models of perfectly competitive markets, with a freely flowing distribution of information and goods from sellers to buyers and no role for investments in marketing. Perhaps more surprising was the explicit omission of brands from the so-called "characteristics approach" to demand modeling during the 1970s. However, that literature focused on objective physical product attributes, dismissing any objective role for marketing other than to identify the seller of a product. In his seminal study of the hedonic framework, Rosen (1974, p. 36) writes: "The terms 'product,' 'model,' 'brand,' and 'design' are used interchangeably to designate commodities of given quality or specification."[1] This oversight is significant, considering the scale of marketing investments. Corrado, Haltiwanger, and Sichel (2005) estimated investment in intangible capital in the US economy, including a broad array of assets such as databases, capitalized research and development, new copyrights and licenses, brand equity, and better organizational structures. In the early 2000s, total investment in intangible capital in the US economy reached about 12 percent of US GDP and nearly one-fifth—roughly $500 billion in 2021—was attributed to marketing expenditures that build and sustain brand equity.

Furthermore, intangible investment is rising as a share of GDP and relative to tangible investment (as discussed in this symposium by Corrado, Haskel, Jona-Lasinio, and Iommi). The share of both total employment and payroll accounted for by occupations that manage brand capital (SOC codes 11-2XXX: sales, marketing, or public relations managers) increased 20–25 percent between 2005 and 2019.

In this article, we discuss many aspects of the concept, measurement, creation, and macro and micro consequences of marketing investments and the intangible capital they create.

Some Facts about Marketing Spending and Intangible Brand Capital

What Is Intangible Brand Capital?

We begin with the concept of a brand. For the purposes of this article, we focus on product brands as opposed to corporate brands, though in many settings the two are synonymous. For instance, Apple has a strong corporate reputation as an

[1] Rosen (1974, p. 37) further adds: "If two brands offer the same bundle, but sell for different prices, consumers only consider the less expensive one, and the identity of sellers is irrelevant to their purchase decisions."

employer, as well as a strong consumer brand reputation. The historic practice of branding consisted primarily of the literal burning of a logo or mark of ownership on a firm's products. However, contemporary marketing experts like Farquar (1989, p. 24) define a brand more broadly as "a name, symbol, design, or mark that enhances the value of a product beyond its functional purpose" where the added value of these enhancements to the basic product are often broadly termed "brand equity."

Brand equity consists of the intangible capital that generates sustainable, incremental profitability to the firm owning the commercial rights to the brand. The expertise, or *human capital*, of the firm's employees in creating and maintaining such brand equity is a related critical economic competence.

Firms brand products through marketing programs that "teach consumers 'who' the product is—by giving it a name and using other brand elements to help identify it—as well as what the product does and why consumers should care" (Keller 2020, p. 38). Branding arises from marketing investments that make consumers aware of the product and persuade them of its benefits and differences relative to the competition. Branding can form associations in the consumer's memory that assist with recall and consideration of the branded product. Branding can also generate perceived differentiation, tangible or intangible, between products.

In this article, we do not discuss the sophisticated strategic steps associated with the design of a brand architecture and its corresponding elements. Instead, we focus on the investments made to communicate and build intangible brand equity. These marketing communication instruments consist of *advertising*, *promotion* (like in-store displays, samples, and merchandising typically near the point of sale), *direct marketing* (including mail, catalogues, and telemarketing), *personal selling* (via the salesforce), *events* (like trade shows), and *public relations* (including media relations, sponsorships, and other mechanisms). According to a survey of chief marketing officers, marketing budgets now represent almost 12 percent of companies' total budgets, on average, up by over 1 percentage point since 2012 (Moorman 2022). In consumer goods industries, marketing budgets regularly approach 25 percent of spending. In our analysis below, we focus primarily on advertising and promotional expenses, as these data are most readily available across firms. There are also good reasons to exclude other potential types of branding investment because direct marketing and personal selling can serve many other non-branding functions, including distribution and pricing.

In the following subsections, we document several recent trends in brand-related investments. In particular, we show that US companies have accelerated their expenditures on advertising, a leading brand-building activity. These investments represent growth in a corresponding aggregate intangible brand capital stock. Over this same period, US firms have also grown their recruiting and payroll shares on the employment of in-house marketing-related personnel.

Advertising and Aggregate Brand Capital

Advertising represents one of the leading instruments for brand investment. According to the most recent IRS Statistics of Income database (Internal Revenue

Service 1954–2018), US corporations expensed $354 billion in advertising spending in 2018, or 1.7 percent of GDP, near the historical average of 1.9 percent. Substantial as it is, this value does not include spending on non-advertising-related brand investments (like public relations, promotional transfers to retailers) nor branding investments made inside the firm (like paying internal employees to design marketing strategies or manage customer accounts). By way of comparison, total tangible nonresidential investment in the national accounts has typically totaled around 13 percent of GDP. Clearly, marketing investments are an important part of firms' efforts to build their capital stocks. Regarding those stocks, we use the IRS data to extend Corrado et al.'s (2016) aggregate advertising-driven brand capital series through 2018. We estimate the total 2018 US brand capital stock to be around $350 billion, more than double the $160 billion estimated (real) stock in 1995. Given that real GDP grew about 75 percent over the same period, advertising-driven brand capital appears to have grown faster than the economy over the past quarter century.

The size of the advertising-driven brand capital stock relative to advertising spending depends on two key assumptions: 1) the capitalization rate for advertising spending and 2) the depreciation rate for advertising capital. The advertising spending capitalization rate consists of the fraction of spending that builds capital that lasts beyond the current period to yield marginal revenue in future periods. The remainder is used up in the current period, and as such represents an expense rather than an investment. Choosing the right capitalization rate is challenging. In practice, firms' brand spending may be a multiple of advertising, capturing other non-advertising sources of marketing, all of which can be incorporated into the capitalization rate. Similarly, the capitalization rate might seek to capture the potential indirect effects of advertising, such as the reinforcing feedback effect of habit formation, brand loyalty, and other persistent responses to advertising. On the other hand, many aspects of a firm's advertising may be transitory, such as the promotion of a temporary price discount or of a promotional product with only temporary distribution. Following Corrado et al. (2016), we assume the capitalization rate is 0.6.

The advertising capital depreciation rate measures the longevity of intangible brand capital stocks. Academics have debated the magnitude of this decay rate since at least the 1960s (for an early survey, see Comanor and Wilson 1979). The debate has by no means been resolved, with some studies finding highly persistent effects and others failing to detect effects lasting more than a few weeks or months. For example, in a cross-section of 55 randomized advertising field experiments for consumer packaged goods (fast-moving pre-packaged, consumer nondurables including food, beverages, health and beauty, and cleaning products), Lodish et al. (1995) not only find that the effects of successful television ad campaign persist more than two years, but the longer-term magnitudes are more than double the immediate-run effects. On the other hand, using a cross-section of 432 digital display-advertising field experiments, Johnson, Lewis, and Nubbemeyer (2017) find that advertising decays at an astonishingly rapid rate of 23 percent per day.

A meta-analysis of older econometric studies finds that 90 percent of the long-run advertising effect (the "duration interval") materializes within 6-9 months (Leone 1995). Taking a longer-term view of a consumer's lifetime brand experiences, Bronnenberg, Dubé, and Gentzkow (2012) estimate an annual brand capital depreciation rate of only 2.5 percent for a cross-section of over 230 product categories of consumer packaged goods.

Perhaps this dispersion of estimates is not surprising. Different forms of advertising may exhibit different degrees of longevity. A short-run price promotion may be forgotten quickly. Yet at the same time it is easy to think of brands—perhaps built through decades of past marketing investments—that now reside virtually permanently in consumers' minds without the need for a lot of explicit repeat prompting. For instance, Coca-Cola was one of the most recalled 2001 Super Bowl ads in the Wall Street Journal-Harris interactive poll, even though Coca-Cola did not broadcast an ad that year (Quick 2001). If we follow Corrado et al.'s (2016) assumption that advertising-driven brand capital depreciates at an annual rate of 55 percent, the calculated advertising capital stock is roughly the same size as current advertising spending in spite of the capitalization ratio and rapid depreciation rate.

Regardless of potential debate around the details, advertising spending and the resulting capital stock are substantial in size. This willingness to expend considerable resources on both the immediate and future effects of advertising indicates that firms perceive such expenditures as valuable. We look at this issue in more detail next.

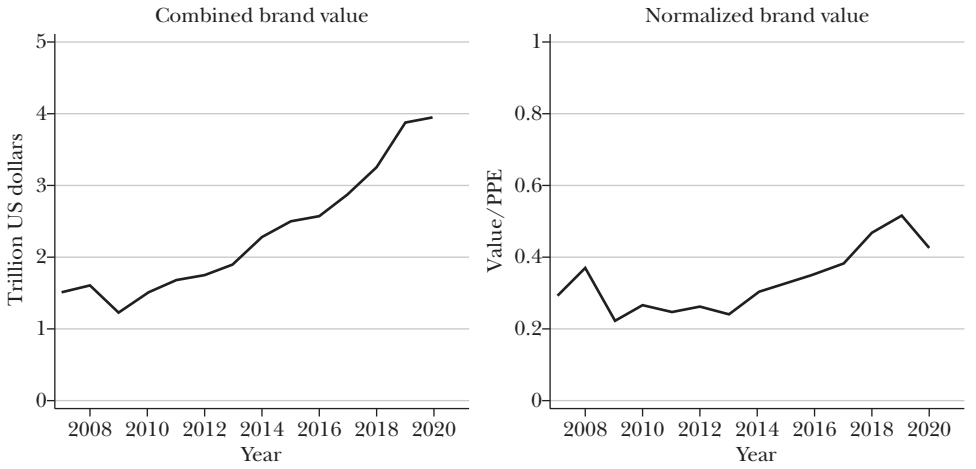
Brand Valuations at the Firm Level

Measuring the full value of brand capital to a firm is notoriously difficult. Conceptually, the value should be defined relative to the counterfactual discounted sum of future profits to the firm *but-for* the commercial rights to the brand and its trademark. This counterfactual raises two challenges. One challenge consists of the appropriate definition of the counterfactual. Do the brand and its trademarked brand elements cease to exist? If so, does the firm build or acquire a different brand instead in the but-for world? Or does the brand get transferred to another (competitor) firm? Another challenge is that, irrespective of the right counterfactual, the but-for profit stream is seldom observed and needs to be imputed or estimated.

Commercial vendors do attempt to compute metrics of brand value, which are widely used by companies in practice. Each vendor uses slightly different methods, but they all involve a mix of forecasts and judgment. We focus on the method used by BrandFinance, a leading brand valuation consultancy, that seeks to estimate the net present value of royalties received from owning a brand, which is close in spirit to the but-for reasoning above.[2]

[2]First, BrandFinance assembles a database of observed industry-specific royalty rates from industry reports. It divides this range into 100 parts. Second, it selects a value from this range, guided by a proprietary brand-strength index, much like a credit rating, that combines the estimated strength, risk, and future potential of a brand relative to its competitors. This brand-strength index is scaled from 0 to 100. If

Figure 1
Brand Value and Normalized Brand Value for the 100 Most Valuable Global Brands



Source: Authors' calculations using data from BrandFinance.
Note: The left panel reports brand value in US dollars. The right panel reports normalized brand value using property, plant, and equipment (PPE).

Figure 1 shows two representations of the value from 2007–2021 of the 100 most valuable brands in the world, as computed by BrandFinance (BrandFinance 2007–2021).[3] The left panel presents the joint value in US dollars, while the right panel normalizes this value by the firms' reported joint value of property, plant, and equipment. The total brand value represented by these 100 most valuable brands is estimated to be $4.14 trillion in 2021 (more than the entire tangible capital stock of Belgium) and has been growing at an average annual growth rate of 8.1 percent. The right panel shows that brand value rose from 29 percent of property, plant, and equipment in 2007–2009 to 47 percent of PPE in 2018–2020. This increase reflects an annual growth rate of 4.8 percent, although this average masks periods of contraction after the Great Recession and at the start of the Covid-19 pandemic.

a brand's score is *X*, the brand specific royalty rate is set to the value of the *X*th increment of the observed royalty rates in the industry to which the brand belongs. Third, to compute the net present value of royalties, BrandFinance's valuation method estimates future revenues from historic revenues, equity analyst forecasts, and economic growth rates, and applies the royalty rate to this forecast. Finally, the post-tax forecasted royalties are discounted to a net present value. For details, see https://brandirectory.com/methodology. For additional information, see ISO Standard 10668 "Brand valuation."

[3]For the sake of comparison, we obtained similar brand valuations for the top 100 global brands from 2007-2021 from another leading vendor, Interbrand (https://interbrand.com/best-global-brands/). While there are some differences, the correlation between the valuations by BrandFinance and Interbrand is 0.88.

Even allowing for considerable measurement error, these numbers are still strikingly large and indicate the importance of brands to the companies that own them. They are also rising faster than GDP and the companies' reported tangible capital. We find similar patterns even if we restrict our attention to the subset of US-based brands.

These measures do depend on a number of strong assumptions, not to mention the fact that we have selected the world's most valuable brands. Nevertheless, other recent research confirms this large value of brand capital, finding that intangible brand capital stocks may represent between 6 and 25 percent of a firm's overall book value using publicly traded US companies (Belo et al. 2022). In a detailed econometric case study of the stacked chips category, Borkovsky et al. (2017) measure the brand value of Pringles at $1.6 billion in 2006, nearly 60 percent of the $2.7 billion for which the Pringles company was sold in 2012. In that study, the brand value was measured relative to a counterfactual market simulation in which Pringles is stripped of its brand equity today, but is permitted to invest in building another brand in the future.

In sum, brand capital stocks constitute an economically large intangible asset to companies. Furthermore, these intangible assets have been growing over time, in spite of mixed findings in the contemporaneous advertising literature regarding the incremental effects of local changes in advertising spending on sales.

Labor and Marketing Expertise as Human Capital

Historically, most companies outsourced the creation of brand capital to consulting firms and advertising agencies. According to a survey of chief marketing officers, approximately one-third of companies' digital marketing is handled by third parties (Moorman 2022). We now document a recent trend of in-sourcing brand marketing and the creation of the highly understudied source of internal marketing expertise, an overlooked source of human capital.

We use the Occupational Employment and Wage Statistics data from the US Bureau of Labor Statistics (Bureau of Labor Statistics 2005–2019), which produces employment and wage estimates annually for nearly 800 occupations each year, to measure corporate investment in internal marketing. Table 1 reports the labor share and payroll share associated with managers who most closely oversee brand capital: sales, marketing, and public relations managers (SOC codes 11-2XXX). We use the years 2005, 2012, and 2019, because occupation codes were reported at a less granular level prior to 2005.

We observe a strong upward trend in marketing personnel both in terms of headcount and payroll share. Payroll share levels are higher, as marketing professionals tend to be white-collar management positions. Most industries experienced double-digit growth rates between 2005 and 2019, with the overall economy experiencing 20 percent growth in marketing managers' labor share and 25 percent growth in their payroll share. These growth rates do not appear to be subsiding, with marketing teams at US firms growing by 12 percent between 2021 and 2022 and 10 percent of firms anticipating continued growth into 2023 (Moorman 2022).

Table 1
Marketing Labor Share of Employment and Payroll

	Employment			*Payroll*		
	2005	*2012*	*2019*	*2005*	*2012*	*2019*
All Economy	**0.44%**	**0.46%**	**0.52%**	**1.12%**	**1.21%**	**1.41%**
Agriculture, Forestry, Fishing and Hunting	0.05%	0.08%	0.09%	0.19%	0.36%	0.33%
Mining, Quarrying, and Oil and Gas Extraction	0.27%	0.32%	0.32%	0.55%	0.69%	0.75%
Utilities	0.48%	0.50%	0.51%	0.87%	0.89%	0.96%
Construction	0.09%	0.10%	0.14%	0.21%	0.23%	0.30%
Wholesale Trade	1.40%	1.56%	1.83%	3.35%	3.84%	4.39%
Information	1.48%	1.69%	2.03%	3.10%	3.56%	4.10%
Finance and Insurance	0.90%	1.02%	1.11%	1.99%	2.35%	2.44%
Real Estate and Rental and Leasing	0.49%	0.47%	0.53%	1.19%	1.32%	1.38%
Professional, Scientific, and Technical Services	0.94%	0.99%	1.49%	1.80%	1.90%	2.77%
Management of Companies and Enterprises	2.76%	3.21%	3.29%	5.09%	5.67%	5.91%
Administrative, Support, and Waste Mgt Services	0.36%	0.26%	0.33%	1.02%	0.80%	1.06%
Educational Services	0.07%	0.11%	0.17%	0.12%	0.21%	0.35%
Health Care and Social Assistance	0.08%	0.08%	0.09%	0.14%	0.15%	0.20%
Arts, Entertainment, and Recreation	0.46%	0.41%	0.42%	1.21%	1.13%	1.24%
Accommodation and Food Services	0.11%	0.09%	0.07%	0.37%	0.33%	0.28%
Other Services (except Public Administration)	0.44%	0.46%	0.57%	1.12%	1.30%	1.69%
Federal, State, and Local Government	0.06%	0.06%	0.07%	0.10%	0.11%	0.11%
Manufacturing	0.57%	0.59%	0.59%	1.50%	1.54%	1.58%
Retail Trade	0.46%	0.51%	0.51%	1.48%	1.59%	1.57%
Transportation and Warehousing	0.22%	0.19%	0.17%	0.47%	0.46%	0.44%

Source: Occupational Employment and Wage Statistics data from the US Bureau of Labor Statistics.
Note: Marketing Labor is defined as sales/marketing/PR managers (occupation codes 11-2XXX).

In addition, our focus on branding communication investments that arise through marketing management excludes the potentially large role in intangible relational capital building of other marketing investments, such as salesforce efforts (relationships between sales reps and their customers) and distribution and retailing efforts (relationships with the *trade*).[4] As one example, salesforce costs represent an additional 5 percent of GDP, or $800 billion (Zoltners, Sinha, and Lorimer 2013), and span over 13 million employees in 2020, close to 10 percent of the US labor force.

These trends coincide with a growing push toward in-sourcing marketing decisions and capabilities. A recent survey by the Association for National Advertisers (2018) finds that 60 percent of US companies have some form of internal marketing, and 78 percent of advertisers have in-house agencies (see also Visser, Sheerin, and Field 2018). These trends suggest a departure from the traditional

[4]The *trade* spans the array of trade partners in the distribution channel between the manufacturer and the end-user consumer, such as wholesalers and retailers. Jointly, these account for 13.7 percent of 2020 US value added and 14.1 percent of European value added (https://unstats.un.org/unsd/snaama/Basic).

model of partnering with advertising agencies outsourcing branding and creative services along with the purchase of advertising media. These trends also appear to be less pronounced in industries where businesses primarily sell to other businesses, where outsourcing of marketing still predominates (as reported by Sweeney 2020).

Brand Capital Investment Theories

Brands would likely exist even in the absence of systematic advertising or other corporate investments in brand-building. After all, consumers frequently rely on a brand's reputation or its trademarked elements, such as logos and colors, to help identify desired products and services. We now discuss established academic theories regarding a firm's private benefits from investments in branding, such as advertising and promotion, that potentially explain the magnitude of economy-wide marketing investments.

Not all advertising and marketing contribute per se to a persistent brand capital stock. For instance, some advertising serves purely to inform consumers about transitory information, as in the case of newspaper feature advertising of a temporary discount at a retail outlet. While these discounts may generate feedback effects—for instance, through brand-buying habits—we focus herein on marketing that contributes directly to persistent brand capital stocks.

We discuss various mechanisms through which marketing investments affect consumer demand and industrial market structure along with the persistence in these effects, reflecting the role of marketing-related intangible capital stocks. We focus on three mechanisms suggested in the literature: 1) reputation and the role of prestige and/or quality; 2) the reduction in transaction and search costs; and 3) competition and the role of strategic interaction and investment escalation. We refer the interested reader to Bronnenberg and Dubé (2017) and Bronnenberg, Dubé, and Moorthy (2019) for more comprehensive discussions of the academic literature on the economics of brands and branding. Additionally, Keller and Aaker (1992) offer rigorous treatments of the perceptual representation of brands in a consumer's memory.

Brand Reputation

Consumers often face incomplete information about a product's quality prior to purchase and consumption. They may prefer branded goods with which they are familiar or that have a reputation for supplying products with certain qualities. In equilibrium, such brand-related reputations can emerge if consumers have a willingness to pay for quality and if a firm with a strong reputation has an incentive to continue to supply high-quality goods in the future to maintain its price premium. As Klein and Leffler (1981, p. 616) write: "[E]conomists also have long considered 'reputations' and brand names to be private devices which provide incentives that assure contract performance in the absence of any third-party enforcer (Hayek 1948, p. 97; Marshall 1949, vol. 4, p. xi)." Whether for packaged goods sold in

supermarkets, retail gasoline or hotels, consumers routinely pay a price premium for branded goods, even when cheaper alternatives are available. When a firm fails to deliver high-quality service, it may even seek to conceal this reputation by changing its name (for example, McDevitt 2011).

Firms with established brands privately benefit from the incremental revenue streams due to 1) high awareness and consideration of their products (for example, Shocker et al. 1991; Laurent, Kapferer, and Roussel 1995) and 2) a reputation for superior *quality* (Bai 2021; McDevitt 2011; McDevitt 2014; Minichilli et al. 2021; Shapiro 1982; Shapiro 1983). These private benefits to the firm can persist over the longer term through brand loyalty, which in turn stems from learning and taste formation (Bronnenberg, Dubé, and Gentzkow 2012; Bronnenberg, Dubé, and Sanders 2020) and from habits and inertia in buying behavior (Keane 1997; Dubé, Hitsch, and Rossi 2009; Dubé, Hitsch, and Rossi 2010). Indeed, many aspects of brand capital are legally protected through the intellectual property rights associated with trademarks, packaging patents, and copyrights.

For present purposes, we focus on a firm's private incentives to invest in the creation and maintenance of a brand through marketing. For instance, the firm may seek to communicate and promote the brand and its reputation to a broader audience for awareness purposes. In some cases, the advertising itself may convey objective information about a product's quality. However, most forms of brand advertising convey little or no objective quality information other than a reminder of the brand. One popular explanation for the prevalence of such uninformative advertising is that the advertising investment itself signals a brand's quality in equilibrium, if high-quality firms derive higher returns from branding than low-quality firms. Similarly, if more efficient firms derive higher returns from branding, consumers may prefer advertised brands because of the signal of higher efficiency and, hence, better deals. However, attempts to test these signaling theories empirically have delivered mixed results, with little evidence of a correlation between product quality and advertising effort. One interesting exception comes from a field experiment for an online restaurant platform that finds the mere disclosure that a restaurant link is a paid ad increases demand for the advertised restaurant (Sahni and Nair 2020).

Another explanation for uninformative advertising is that consumers derive consumption utility from the brand itself. According to this *persuasion* or *prestige* view, marketing expenditure in advertising and other forms of branding can create a consumable intangible service (say, prestige or lifestyle) that is complementary to the branded good or service (Becker and Murphy 1993). For instance, Kamenica, Naclerio, and Malani (2013) find that exposure to advertising for a branded antihistamine causes an increase in the rate at which the drug works—a physiological advertising effect.

A more cynical view of uninformative advertising is that it persuades consumers to perceive spurious differentiation between products, potentially causing spurious sources of loyalty (and for the sellers, market power). For example, branded headache medicines generate higher total revenues and typically sell at a significant price premium over objectively identical store brands that differ only in terms of

brand name and branding elements. Meanwhile, pharmacists and physicians are considerably more likely to choose store-brand headache medicines than socio-demographically similar consumers who lack the healthcare domain expertise to realize the lack of objective differentiation (Bronnenberg et al. 2015).

Reductions in Consumer Transaction Costs and Search Frictions

In many shopping contexts, consumers incur transaction costs prior to making a decision. These costs can be internal like thinking and deliberation or external like browsing and research. They may additionally include negotiation, ordering and payment, delivery, and post-purchase service and support. These transaction-related costs can consume both time and money. According to the 2019 US time-use survey, consumers spend 0.75 hours per day purchasing goods and services on average, which corresponds to 1.71 hours per day for those who do any purchasing at all.[5] Similarly, the empirical literature on consumer search has routinely estimated large search costs (for example, Honka 2014; Kim et al. 2010).

Consumers may choose branded goods because they are less costly to consider and evaluate. The ability to recognize a brand and recall associated product information about the branded good from memory can help a consumer avoid several of these transaction costs. This information could include quality, product attributes, or the likely price being charged. It may be triggered through recall and memory if, for instance, branding helps a consumer recall past experiences with a branded good. Alternatively, this information may be conveyed directly through the branding elements. For instance, the strong effect of tobacco packaging color on consumers' perceptions of the quality of the tobacco led Australia to implement a "plain packaging" regulation requiring all sellers to adopt a common, drab-brown packaging color.

Thus, investments in brand advertising can generate a persistent reduction in transaction costs by increasing the prominence of a brand in a consumer's memory, or making it "top-of-mind." For instance, advertising has been found to increase the likelihood of being considered by consumers at the point of sale (for example, Draganska and Klapper 2011). Consumers are also more likely to direct their search to more prominently branded retailers (for example, Baye, De los Santos, and Wildenbeest 2016) and may be more likely to click on firms with more prominent positions in search results on an online platform (for example, Ursu 2018). In principle, the long-term effects of branding on transaction costs could be self-reinforcing if consumers are more likely to consider and purchase branded goods, thereby establishing persistent *consumption capital* (or "habits") for those goods.

Competition and Equilibrium Brand Investment

Thus far, we have discussed a firm's incentives to invest in branding from the perspective of the monetizable equity a brand can create for consumers and

[5] These estimates are from the US Bureau of Labor Statistics. See https://www.bls.gov/tus/a1-2019.pdf.

demand. We now turn to equilibrium theories of branding and the strategic incentives on the supply side for brand investments. In particular, strategic considerations can either stimulate or deter branding efforts.

The strategic incentives for branding depend on the nature of marketing productivity and the returns to branding. Constant (or even increasing) returns to branding that sustain a high marginal impact of these investments, even at high levels of investment, can lead to an escalation in advertising or other forms of marketing in equilibrium.

We start with the assumption of constant returns to scale in branding. In the special case where the impact of branding expenditures on demand for the branded good does not affect the own-price elasticity, we obtain a classic result: the optimal advertising-to-sales ratio equals the ratio of the advertising elasticity to the own-price elasticity (Dorfman and Steiner 1954). One positive implication of this result is that firms in more competitive markets have less incentive to invest in branding. However, this prediction hinges on the assumption that advertising does not affect the price elasticity of demand.

Next, consider the case of economies of scale in branding and the potential for escalation. Suppose branding expenditures are endogenously chosen by the firm, but are fixed and sunk, and create brand capital that increases future demand. The strategic interaction of firms in this setting can lead to an escalation in marketing investments that creates barriers to entry, sustaining market power and concentration (Sutton 1991). Even as the market becomes very large, an escalation in brand spending arises without a corresponding escalation in entry, so that only a small number of branded goods dominate while charging a price premium. The escalation in advertising may be even higher if early entrants use their branding to preempt future entry by a rival.

Researchers have documented such outcomes extensively in the consumer packaged goods industry. The typical category within consumer packaged goods has been dominated by the same small set of established brands for decades, with early (surviving) entrants typically sustaining a higher share than later entrants (for example, Sutton 1991; Bronnenberg, Dhar, and Dubé 2011).

Interestingly, the rapid shift away from traditional television advertising to increasingly targetable and personalizable digital advertising could potentially upend the market structure of consumer goods industries. Most television advertising is purchased upfront, months before the airing of the ad and the sale of the product. On the other hand, digital ads are typically targeted to individual consumers contemporaneously as they browse and evolve towards the purchase decision (the so-called *purchase funnel*). According to a survey of chief marketing officers, digital marketing now accounts for 57 percent of marketing budgets (Moorman 2022). In addition, whereas television advertising is mostly borne as a fixed and sunk cost, digital advertising is typically borne as a marginal cost, which can theoretically lead to fragmentation with a large number of small (low-advertising) brands. During the past decade, many categories of consumer packaged goods have begun to fragment, as new local *craft* brands have begun to steal share from established brands. The beer

industry is an oft-cited example (for example, Elzinga 2011; Bronnenberg, Dubé, and Joo forthcoming).

When advertising is primarily *combative*, it primarily shifts share from one competitor to another in a tug-of-war. Firms may find themselves in a prisoner's dilemma in which all firms would prefer to cooperate to reduce overall advertising spending, but they are in a non-cooperative outcome in which each must advertise to defend against rivals' advertising. In some instances, firms may see no net change in their market shares in equilibrium despite large advertising outlays. Such prisoner's dilemmas have been documented in both laboratory settings (for example, Corfman and Lehmann 1994; Chen et al. 2009) and in the over-the-counter market for painkillers (for example, Anderson et al. 2016).

In contrast, market forces may deter firms from investing in branding when there are positive externalities on other firms. For instance, advertising by one firm may increase awareness for the entire category, generating positive spillovers to rivals. In this case, firms may free-ride off one-another's brand capital without internalizing the benefits their advertising generates for rivals. Such spillovers have been documented empirically in the market for antidepressants (Shapiro 2018), statins (Sinkinson and Starc 2019), and digital platforms for restaurant delivery (Sahni 2016). Shapiro (2018) finds that advertising would increase 50 percent in the antidepressants market if firms hypothetically cooperated on their advertising, so that each firm both paid its "share" of the industry advertising and knew that other firms were doing so as well.

Marketing and Social Welfare

Here, we turn to the divisive debate regarding the social benefits of brands and brand investments. While most of the debate has focused on advertising, the incentives for advertising are not distinct from the incentives to invest in other communication strategies to build brands.

The Persuasive View

In the *persuasive view*, advertising conveys information from an "interested" party, thereby providing little objective value and mostly creating spurious perceived differentiation and loyalty (for example, Marshall 1919; Kaldor 1950; Galbraith 1958; Solow 1967). Indeed, consumers are often empirically unable to identify their preferred brands in blind taste tests (Husband and Godfrey 1934; Thumin 1962; Allison and Uhl 1964) and in some instances prefer a cheaper store brand (as in Bronnenberg, Dubé, and Sanders 2020). Furthermore, such persuasive advertising can generate barriers to entry that sustain high prices and reputational monopolies. Economies of scale in branding would bolster these barriers to entry. In short, under the persuasive view, advertising is necessarily excessive because it decreases welfare, facilitating higher prices with no objective increase in consumer utility.

As noted earlier, established advertised brands have persistently dominated markets for consumer packaged goods for at least half a century, with the earliest entrants out-performing later entrants (Bronnenberg, Dhar, and Dubé 2009). Similarly, equilibrium advertising levels escalate in larger geographic markets, with no corresponding increase in the number of branded competitors (Bronnenberg, Dhar, and Dubé 2007). Some of this advertising could be socially wasteful. According to a Food and Drug Administration (2022) information website, generic prescription drugs are typically 80–85 percent cheaper than the equivalent branded drug. Overall, the Food and Drug Administration (2016) summarizes evidence that patients could reduce their daily drug costs by 14–16 percent if they switched to generics, which corresponds to an economy-wide annual saving of $17 billion. Bronnenberg et al. (2015) estimate that consumers could save $44 billion annually simply by switching to store-branded consumer packaged goods. However, weaning consumers off premium-priced branded goods is often difficult even when a cheaper, physically comparable alternative is available. The provision of objective information about the comparability of the cheaper variant may have be sufficient to switch consumers away from the established brand (for example, Cox, Coney, and Ruppe 1983; Carrera and Villas-Boas 2015; Bronnenberg, Dubé, and Sanders 2020). Furthermore, if one views such effects as reflecting in part the complementary view of advertising discussed below, welfare interpretations become more difficult.

The Informative View

During the 1960s, a competing *informative view* of advertising emerged, led primarily by Chicago-school economists (Stigler 1961; Telser 1964). Advocates argued that advertising communicates valuable information about the product and its attributes. To the extent that advertising reduces consumer search and evaluation costs, it would be procompetitive, leading to less price dispersion and lower markups. Furthermore, advertising could facilitate entry and further toughen competition. Under the informative view, advertising can be socially beneficial by creating consumer value and making markets more competitive.

Some evidence supports the welfare-improving potential of advertising. For instance, antidepressant advertising has been found to increase prescriptions and most striking, to decrease workplace absenteeism (Shapiro 2020). Similarly, advertising during US presidential elections may have a large effect on voter turnout, stimulating political participation (Shachar 2009; Gordon and Hartmann 2013). Advertising by branded incumbents for cholesterol-reducing statins has been found to facilitate entry by unbranded generic competitors (Sinkinson and Starc 2019). As discussed above, the branding literature finds that the mental associations created by brands in a consumer's memory help reduce search and deliberation costs at the point of sale.

Some reputational benefits of branding may be welfare-improving, too. Bai (2021) finds that introducing a branding technology in the market for watermelons quickly led to higher quality in equilibrium. Similarly, biosimilar branded

and branded generic drugs in Chile were found to be of much higher quality than cheaper unbranded alternatives (Atal, Cuesta, and Sæthre 2019). In that setting, regulations that limited entry of low-quality biosimilars increased consumer welfare, in spite of leading to higher prices.

As we noted, advertising can sometimes increase overall consumer interest in the product category, generating potential spillovers between firms. In practice, the free-riding problem can lead to under-investment relative to the social optimum for such class-expanding advertising.

The Complementary View

A more recent stream of literature takes the *complementary* view of advertising, whereby the consumer derives consumption utility from the brand and branding itself, even if the advertising conveys no objective information (Becker and Murphy 1993). Empirically, consumers who have recently purchased a branded good are more likely to watch (consume) ads for that good instead of skipping them (Tuchman 2019). A similar complementarity was documented between advertising during the National Football League's Super Bowl for a given brand and subsequent consumption of that brand during future sporting events (Hartmann and Klapper 2018).

The welfare implications of advertising are more ambiguous under the complementary view, which treats advertising as a consumption good in and of itself. However, Becker and Murphy (1993) show that if advertising decreases the equilibrium price of the advertised good, then the market is under-supplying advertising. Intuitively, this test would indicate that firms are not taking into account advertising's ability to increase willingness-to-pay for the advertised good when deciding their marketing spending. Conversely, even if advertising increases equilibrium prices, it need not be socially excessive as long as it creates enough consumer value.

A Roadmap for Future Research

We see at least three potentially valuable directions for future research on the economics of brand capital.

Agency and Conflict of Interest

Many firms rely on external advertising agencies not just to buy and allocate advertising media, but also to evaluate the performance of the ads. This joint duty of purchasing and auditing the performance of advertising raises a clear conflict of interest. Other firms assign a marketing budget to an internal team to conduct the media buying and performance evaluation. Again, there is a conflict of interest. In either case, even if those in charge of the marketing budget do not literally obfuscate negative evidence, they face little incentive to seek out more reliable methods.

As an example of the potential for such conflicts of interest, Blake, Nosko, and Tadelis (2015) show how simple ordinary least squares estimation suggests the presence of strong and significant effects for the eBay company to engage in *paid search* advertising so that its name would appear at the top of a search engine result, implying a return-on-investment of over 1,000 percent. If a marketing team that was hired to purchase such advertising found such evidence, it would have little incentive to assess its robustness, even though it is based only on a correlation. Blake, Nosko, and Tadelis (2015) then develop more reliable experimental evidence that paid brand-keyword search advertising at eBay had a very small effect on demand, because over 95 percent of that effect consisted of cannibalization of traffic to eBay that would have come free through the organic channel. Essentially, eBay had been paying search engines to place their site at the top of the search list when browsers searched for "ebay," even though the site would have certainly also been at the top of the list of "organic" (not-paid-for) search results. The true return-on-investment was approximately –75 percent, and eBay subsequently terminated its brand keyword search campaigns on which it had invested $30 million in 2010 alone. In a follow-up study of branded keyword search advertising, Simonov, Nosko, and Rao (2018) find a similar cannibalization effect—that is, paid advertising was just redirecting traffic that would have arrived without this advertising—by conducting randomized field experiments for the 2,500 most searched brands on the Bing search engine.

The built-in conflict of interest in making and evaluating decisions about advertising would be expected to lead a wide variation in the outcomes of advertising, and indeed, the long empirical literature measuring the effect of advertising has routinely documented mixed results. Aaker and Carmen (1982) speculate that some of these mixed findings reflect a tendency for established brands to over-invest in advertising, with some of the budget spent on wasteful and ineffective branding. For example, one might expect to find little or no effect of local changes in advertising for established brands already in possession of large, intangible brand capital stocks. Indeed, Shapiro, Hitsch, and Tuchman (2021) find small and mostly insignificant advertising effects for almost 300 of the top-advertised consumer brands. In contrast, using randomized television advertising experiments, Lodish et al. (1995) find much larger television advertising effects for new products, often persisting several years after a campaign. Given the long-standing expertise of producers of consumer packaged goods in advertising, it is surprising to find widespread investment in ineffective advertising—unless one takes seriously the agency problems in advertising spending. There are of course examples of large advertising effects, including for established brands. For instance, advertising during the football Super Bowl, one of the most expensive and controversial forms of advertising, has been found to increase sales for branded consumer goods and for movies (Stephens-Davidowitz, Varian, and Smith 2017; Hartmann and Klapper 2018).

The mixed results are not merely an artifact of the consumer packaged goods industry. For example, Shapiro (2018) finds a precise null effect of advertising for health insurance. Of course, these studies raise some questions about the appropriate capitalization rate to assign to advertising spending.

Productivity and Growth

In different ways, brand capital can cause firm-level measures of productivity to be either overstated or understated.

In practice, the role of brand capital is almost inevitably unmeasured in production analysis. The standard productivity measures that are constructed for a firm use only tangible inputs and outputs. Such measures will capture the output that the intangible creates, but they do not count brand capital among the inputs. As a result, companies with a large amount of brand capital (and/or a high elasticity of output to brand) will appear to have high measured productivity, although the firm's true (intangible-adjusted) productivity level would be much lower. Given the enormous variations in measured productivity among firms even in narrowly defined markets, it is possible—and in some markets probable—that some of this variation is coming from differences in the size or efficacy of firms' brands.

One nuance here resides in how output is measured. As noted, brand does not lead directly to more physical output per unit of input (or for service-producing firms, more countable units of anything). Instead, it raises the prices at which those units are sold. Thus, quantity-based measures of productivity will not capture the effect of brand on output, while revenue-based measures will. De Loecker and Syverson (2021) provide a broader discussion of the respective strengths and weaknesses of quantity- and revenue-based productivity measures.

The discussion to this point takes brand capital as installed and considers its effect on production and measured productivity. However, when brand investments are being made, they are (conceptually) an output of the firm, as would be the case for a firm producing a tangible investment good. In this way, investments in brand capital cause productivity to be understated. The firm looks like it is employing many resources without obtaining a lot of output from them, but in reality that output is not being counted. Thus, when brand capital is first produced it causes measured productivity to *understate* the true productivity level of the firm.

The net effect on measured productivity of the overstatement due to not measuring brand capital as an input and the overstatement from not measuring brand capital creation as an output depends on the relative size and timing of firms' brand investments and installed stocks. Brynjolfsson, Rock, and Syverson (2021) discuss this and related effects of intangible capital on productivity measurement more generally.

Alternative Sources of Intangible Brand Capital

Our discussion has focused on communication investments to build and maintain intangible brand capital. The practice of marketing is, of course, broader in scope; in particular, it includes non-branding investments.

As one example, firms invest in "customer relationship management" systems, which both seek to acquire new customers and to sell more to existing customers through upselling and cross-selling. The tools of customer relationship management take the form of incentives, convenient transactions, and information about what the firm offers. Foster, Haltiwanger, and Syverson (2016) are among the first

to study how demand-side fundamentals, such as multi-year efforts to build a customer base (and create customer relationships), explain the slow growth of new plants in commodities industries. The properties of such a customer base, or "demand stock" (p. 97), and how it is affected by investments in marketing, remain an open question.

In addition, consumption itself can be an important source of intangible capital and manufacturer/consumer relations. Consumers form preferences for products they have consumed in the past (for example, Bronnenberg, Dubé, and Gentzkow 2012; Atkin 2013), and for products consumed by their parents (Anderson et al. 2016). Such consumption capital can lead to the formation of preferences that bring important advantages to firms (Bain 1956). However, not much is known about the moderating effect of marketing investments on the formation of consumption capital throughout a consumer's lifetime. There are some suggestive examples. Bronnenberg, Dubé, and Joo (forthcoming) link current consumer preferences for craft beers to historical local availability and distribution, while Atkin (2013) documents that past prices of staple foods impact current preferences. However, more study is needed of the way in which non-branding investments in marketing may initiate and help form long-lived relations between firms and their customers.

Conclusion

The economics literature in recent decades has largely overlooked the role of branding and marketing human capital for our understanding of markets and their organization as well as firm productivity and macroeconomic growth. We acknowledge the potential for some branding efforts to be socially wasteful. However, we also see ample scope for a welfare-improving role of brand capital through its ability to facilitate consumer search and evaluation. Furthermore, we see reasonable potential for brands and branding to offer more than transaction services; in some instances, they create genuine consumption benefits. Given the large economic magnitude of intangible brand capital and its recent growth, these issues seem likely to be of first-order importance.

■ *We are very grateful to Lia Kim for excellent research assistance. Dubé acknowledges research support from the Kilts Center for Marketing and the Charles E. Merrill faculty research fund.*

References

Aaker, D.A. 1982. "A Study of Advertising Generalization." *Journal of Marketing* 44 (1): 10–18.

Aaker, K. L. 1992. "The Effects of Sequential Introduction of Brand Extensions." *Journal of Marketing Research* 29 (1): 35–50.

Allison, Ralph I., and Kenneth P. Uhl. 1964. "Influence of Beer Brand Identification on Taste Perception." *Journal of Marketing Research* 1 (3): 36–39.

Anderson, Simon P., Federico Ciliberto, Jura Liaukonyte, and Régis Renault. 2016. "Push-Me Pull-You: Comparative Advertising in the OTC Analgesics Industry." *Rand* 47 (4): 1029–56.

Association of National Advertisers. 2018. *The Continued Rise of the In-House Agency*. New York: Association of National Advertisers.

Atal, Juan Pablo, Jose Ignacio Cuesta, and Morten Sæthre. 2019. "Quality Regulation and Competition: Evidence from Pharmaceutical Markets." Penn Institute for Economic Research Working Paper 19–017.

Atkin, D. 2013. "Trade, Tastes, and Nutrition in India." *American Economic Review* 103 (5): 1629–63.

Bai, Jie. 2021. "Melons as Lemons: Asymmetric Information, Consumer Learning and Seller Reputation." CID Faculty Working Paper 396.

Bain, Joe. S. 1956. *Barriers to New Competition*. Cambridge, MA: Harvard University Press.

Baye, Michael R., Babur De los Santos, and Matthijs R. Wildenbeest. 2016. "Search Engine Optimization: What Drives Organic Traffic to Retail Sites?" *Journal of Economics and Management Strategy* 25 (1): 6–31.

Becker, Gary, and Kevin M. Murphy. 1993. "A Simple Theory of Advertising as a Good or Bad." *Quarterly Journal of Economics* 108 (4): 941–64.

Belo, Frederico, Vito D. Gala, Juliana Salomao, and Maria Ana Vitorino. 2022. "Decomposing Firm Value." *Journal of Financial Economics* 143 (2): 619–39.

Blake, Thomas, Chris Nosko, and Steven Tadelis. 2015. "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment." *Econometrica* 83 (1): 155–74.

Borkovsky, Ron N., Avi Goldfarb, Avery M. Haviv, and Sridhar Moorthy. 2017. "Measuring and Understanding Brand Value in a Dynamic Model of Brand Management." *Marketing Science* 36 (4): 471–99.

Braithwaite, Dorothea. 1928. "The Economic Effects of Advertisement." *Economic Journal* 38 (149): 16–37.

BrandFinance. 2007–2021. "Global Rankings: Brand Rankings by Sector and Country." https://brandirectory.com/rankings (accessed December 15, 2021).

Bronnenberg, Bart J., Sanjay K. Dhar, and Jean-Pierre Dubé. 2007. "Consumer Packaged Goods in the United States: National Brands, Local Branding." *Journal of Marketing Research* 44 (1): 4–13.

Bronnenberg, Bart J., Sanjay K. Dhar, and Jean-Pierre Dubé. 2009. "Brand History, Geography, and the Persistence of Brand Shares." *Journal of Political Economy* 117 (1): 87–115.

Bronnenberg, Bart J., Sanjay K. Dhar, and Jean-Pierre Dubé. 2011. "Endogenous Sunk Costs and the Geographic Differences in the Market Structures of CPG Categories." *Quantitative Marketing and Economics* 9: 1–23.

Bronnenberg, Bart J., Jean-Pierre Dubé, and Matthew Gentzkow. 2012. "The Evolution of Brand Preferences: Evidence from Consumer Migration." *American Economic Review* 102 (6): 2472–508.

Bronnenberg, Bart J., Jean-Pierre Dubé, and Sridhar Moorthy. 2019. "The Economics of Brands and Branding." In *Handbook of the Economics of Marketing*, edited by Jean-Pierre Dubé and Peter Rossi, 291–358. Amsterdam: Elsevier.

Bronnenberg, Bart J., Jean-Pierre Dubé, and Robert E. Sanders. 2020. "Consumer Misinformation and the Brand Premium: A Private Label Blind Taste Test." *Marketing Science* 39 (2), 382–406.

Bronnenberg, Bart J., Jean-Pierre Dubé, and Chad Syverson. 2022. "Replication data for: Marketing Investment and Intangible Brand Capital." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E168201V1.

Bronnenberg, Bart J., Jean-Pierre Dubé, Matthew Gentzkow, and Jesse M. Shapiro. 2015. "Do Pharmacists Buy Bayer? Sophisticated Shoppers and the Brand Premium." *Quarterly Journal of Economics* 130 (4): 1669–726.

Bronnenberg, Bart J., and Jean-Pierre Dubé. 2017. "The Formation of Consumer Brand Preferences." *Annual Review of Economics* 9: 353–82.

Bronnenberg, Bart J., Jean-Pierre Dubé, and Joonhwi Joo. Forthcoming. "Millennials and the Take-Off of

Craft Brands: Preference Formation in the U.S. Beer Industry." *Marketing Science*.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2021. "The Productivity J-Curve: Intangibles Complement General Purpose Technologies." *American Economic Journal: Macroeconomics* 13 (1): 333–72.

Bureau of Economic Analysis. 1970–2020. "National Income and Products Accounts, Percentage Shares of Gross Domestic Product." US Department of Commerce. https://apps.bea.gov/iTable/iTable.cfm?reqid=19&step=2#reqid=19&step=2&isuri=1&1921=survey (accessed May 4, 2022).

Bureau of Labor Statistics. 2005–2019. "Occupational Employment and Wage Statistics." US Department of Labor. https://www.bls.gov/oes/tables.htm (accessed April 25, 2022).

Carrera, M., and S. B. Villas-Boas. 2015. "Generic Aversion and Observational Learning in the Over-the-Counter Drug Market." Unpublished.

Chen, Yuxin, Yogesh V. Joshi, Jagmohan S. Raju, and Z. John Zhang. 2009. "A Theory of Combative Advertising." *Marketing Science* 28 (1): 1–19.

Coase, R. H. 1937. "The Nature of the Firm." *Economica* 4 (16): 386–405.

Comanor, William S., and Thomas A. Wilson. 1979. "The Effect of Advertising on Competition: A Survey." *Journal of Economic Literature* 17 (2): 453–76.

Corfman, Kim P., and Donald R. Lehmann. 1994. "The Prisoner's Dilemma and the Role of Information in Setting Advertising Budgets." *Journal of Advertising* 23 (2): 35–48.

Corrado, Carol A., and Janet X. Hao. 2013. "Brands As Productive Assets: Concepts, Measurement, and Global Trends." World Intellectual Property Organization (WIPO) Economics and Statistics Working Paper 13.

Corrado, Carol, John Haltiwanger, and Dan Sichel. 2005. *Measuring Capital in the New Economy*. Chicago: University of Chicago Press.

Corrado, Carol, Jonathan Haskel, Massimiliano Iommi, and Cecilia Jona-Lasinio. 2020. "Intangible Capital, Innovation, and Productivity à la Jorgenson: Evidence from Europe and the United States." In *Measuring Economic Growth and Productivity: Foundations, KLEMS Production Models, and Extensions*, edited by B. M. Fraumeni, 363–85. London: Academic Press.

Corrado, Carol, Jonathan Haskel, Cecilia Jona-Lasinio, and Massimiliano Iommi. 2016. "Intangible Investment in the EU and US before and since the Great Recession and Its Contribution to Productivity Growth." European Investment Bank Working Paper 2016/08.

Cox, Steven R., Kenneth A. Coney, and Peter F. Ruppe. 1983. "The Impact of Comparative Product Ingredient Information." *Journal of Public Policy & Marketing* 2 (1): 57–69.

de Loecker, Jan, and Chad Syverson. 2021. "An Industrial Organization Perspective on Productivity." In *Handbook of Industrial Organization*, Vol. 4, edited by Kate Ho, Ali Hortacsu, and Alessandro Lizzeri, 141–223. Amsterdam: Elsevier.

de Rassenfosse, Gaétan, and Adam B. Jaffe. 2017. "Econometric Evidence on the R&D Depreciation Rate." NBER Working Paper 23072.

Dixit, Avinash K., and Joseph E. Stiglitz. 1977. "Monopolistic Competition and Optimum Product Diversity." *American Economic Review* 67 (3): 297–308.

Draganska, Michaela, and Daniel Klapper. 2011. "Choice Set Heterogeneity and the Role of Advertising: An Analysis with Micro and Macro Data." *Journal of Marketing Research* 48 (4): 653–69.

Dubé, Jean-Pierre, Günter J. Hitsch, and Peter E. Rossi. 2009. "Do Switching Costs Make Markets Less Competitive?" *Journal of Marketing Research* 46 (4): 435–45.

Dubé, Jean-Pierre, Günter J. Hitsch, and Peter E. Rossi. 2010. "State Dependence and Alternative Explanations for Consumer Inertia." *RAND Journal of Economics* 41 (3): 417–45.

Elzinga, Kenneth G. 2011. "The U.S. Beer Industry: Concentration, Fragmentation, and a Nexus with Wine." *Journal of Wine Economics* 6 (2): 217–30.

Farquhar, Peter H. 1989. "Managing Brand Equity." *Marketing Research* 1 (3): 24–33.

Food and Drug Administration. 2016. "Savings From Generic Drugs Purchased at Retail Pharmacies." US Food and Drug Administration. https://www.fda.gov/drugs/resources-you-drugs/savings-generic-drugs-purchased-retail-pharmacies (accessed on July 6, 2022).

Food and Drug Administration. 2022. "Generic Drugs: Questions & Answers." US Food and Drug Administration. https://www.fda.gov/drugs/questions-answers/generic-drugs-questions-answers#q4 (accessed on June 6, 2022).

Foster, Lucia, John Haltiwanger, and Chad Syverson. 2016. "The Slow Growth of New Plants: Learning about Demand?" *Economica* 83 (329): 91–129

Galbraith, J.K. 1958. *The Affluent Society*. Boston: Houghton Mifflin.

Galbraith, J.K., and John D. Black. 1935. "The Quantitative Position of Marketing in the United States." *Quarterly Journal of Economics* 49 (3): 394–413.

Gordon, Brett R., and Wesley R. Hartmann. 2013. "Advertising Effects in Presidential Elections." *Marketing Science* 32 (1): 1–189.

Hartmann, Wesley R., and Daniel Klapper. 2018. "Super Bowl Ads." *Marketing Science* 37 (1): 78–96.

Honka, Elisabeth. 2014. "Quantifying Search and Switching Costs in the US Auto Insurance Industry." *RAND Journal of Economics* 45 (4): 847–84.

Husband, R.W., and Godfrey. 1934. "An Experimental Study of Cigarette Identification." *Journal of Applied Psychology* 18: 220–23.

Internal Revenue Service. 1954–2018. "Statement of Income Data." Internal Revenue Service. https://www.irs.gov/statistics/soi-tax-stats-corporation-income-tax-returns-complete-report-publication-16 (accessed January 4, 2022).

Johnson, Garrett A., Randall A. Lewis, and Elmar I. Nubbemeyer. 2017. "Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness." *Journal of Marketing Research* 54 (6): 867–84.

Kaldor, Nicholas. 1950. "The Economic Aspects of Advertising." *Review of Economic Studies* 18 (1): 1–27.

Kamenica, Emir, Robert Naclerio, and Anup Malani. 2013. "Advertisements Impact the Physiological Efficacy of a Branded Drug." *Proceedings of the National Academy of Sciences* 110 (32): 12931–5.

Keane, Michael. P. 1997. "Modeling Heterogeneity and State Dependence in Consumer Choice Behavior." *Journal of Business & Economic Statistics* 15 (3): 310–27.

Keller, Kevin Lane, and David A. Aaker. 1992. "The Effects of Sequential Introduction of Brand Extensions." Journal of Marketing Research 29 (1): 35-50.

Keller, Kevin Lane, and Vanitha Swaminathan. 2020. *Strategic Brand Management*. Harlow: Pearson Education Limited.

Kim, Jun B. 2010. "Online Demand Under Limited Consumer Search." *Marketing Science* 29 (6): 963–1169.

Klein, Benjamin, and Keith B. Leffler. 1981. "The Role of Market Forces in Assuring Contractual." *Journal of Political Economy* 89 (4): 615–41.

Laurent, Gilles, Jean-Noel Kapferer, and Françoise Roussel. 1995. "The Underlying Structure of Brand Awareness Scores." *Marketing Science* 14 (3): G170–9.

Leone, Robert P. 1995. "Generalizing What is Known about Temporal Aggregation and Advertising Carryover." *Marketing Science* 14 G141–50.

Liang, Ting-Pend, and Jin-Shiang Huang. 1998. "An Empirical Study on Consumer Acceptance of Products in Electronic Markets: a Transaction Cost Model." *Decision Support Systems* 24 (1): 29–43.

Lodish, Leonard M., Magid M. Abraham, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, and Mary Ellen Stevens. 1995. "A Summary of Fifty-Five In-Market Experimental Estimates of the Long-Term Effect of TV Advertising." *Marketing Science* 14 (3): G133–40.

Marshall, Alfred. 1919. *Industry and Trade*. London: Macmillan and Co.

McDevitt, Ryan Carty. 2011. "Names and Reputations: An Empirical Analysis." *American Economic Journal: Microeconomics* 3: 193–209.

McDevitt, Ryan C. 2014. "'A' Business by Any Other Name: Firm Name Choice as a Signal of Firm Quality." *Journal of Political Economy* 122 (4): 909–44.

Minichilli, Alessandro, Annalisa Prencipe, Suresh Radhakrishnan, and Gianfranco Siciliano. 2021. "What's in a Name? Eponymous Private Firms and Financial Reporting Quality." *Management Science* 68 (3): 1591–2376.

Moorman, Christine. 2022. *The CMO Survey: The Highlights and Insights Report*. Durham: CMO Survey.

Quick, Rebecca. 2001. "Interesting Ads Held Viewers That the Super Bowl Bored." *Wall Street Journal*, February 2. https://www.wsj.com/articles/SB981081015134599674.

Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy* 82 (1): 34–55.

Sahni, Navdeep S. 2016. "Advertising Spillovers: Evidence from Online Field Experiments and Implications for Returns on Advertising." *Journal of Marketing Research* 53 (4): 459–78.

Sahni, Navdeep S., and Harikesh S. Nair. 2020. "Does Advertising Serve as a Signal? Evidence from a Field Experiment in Mobile Search." *Review of Economic Studies* 87 (3): 1529–1564.

Shachar, Ron. 2009. "The Political Participation Puzzle and Marketing." *Journal of Marketing Research* 46 (6): 798–815.

Shapiro, Bradley T. 2018. "Positive Spillovers and Free Riding in Advertising of Prescription Pharmaceuticals: The Case of Antidepressants." *Journal of Political Economy* 126 (1): 381–437.

Shapiro, Bradley T. 2020. "Promoting Wellness or Waste? Evidence from Antidepressant Advertising."

forthcoming *American Economic Journal: Microeconomics* 14 (2): 439–77.

Shapiro, Bradley T., Günter J. Hitsch, and Anna E. Tuchman. 2021. "TV Advertising Effectiveness and Profitability: Generalizable Results from 288 Brands." Econometrica 89 (4): 1855–79.

Shapiro, Carl. 1982. "Consumer Information, Product Quality, and Seller Reputation." *Bell Journal of Economics* 13 (1): 20–35.

Shapiro, C. 1983. "Premiums for High Quality Products as Returns to Reputations." *Quarterly Journal of Economics* 98: 659–79. Retrieved from http://www.jstor.org/stable/1881782.

Shaw, A.W. 1912. "Some Problems in Market Distribution." *Quarterly Journal of Economics* 26 (4): 703–65.

Shocker, Allan D., Moshe Ben-Akiva, Bruno Boccara, and Prakash Nedungadi. 1991. "Consideration Set Influences on Consumer Decision-Making and Choice: Issues, Models, and Suggestions." *Marketing Letters* 2 (3): 181–97.

Simonov, Andrey, Chris Nosko, and Justin M. Rao. 2018. "Competition and Crowd-Out for Brand Keywords in Sponsored Search." *Marketing Science* 37 (2): 200–15.

Sinkinson, Michael, and Amanda Starc. 2019. "Ask Your Doctor? Direct-to-Consumer Advertising of Pharmaceuticals." *Review of Economic Studies* 86 (2): 836–81.

Solow, Robert M. 1967. The New Industrial State or Son of Affluence. *Public Interest* 9: 100–8.

Stephens-Davidowitz, Seth, Hal Varian, and Michael D. Smith. 2017. "Super Returns to Super Bowl Ads?" *Quantitative Marketing and Economics* 15 (1): 1–28.

Stigler, George J. 1961. "The Economics of Information." *Journal of Political Economy* 69 (3): 213–25

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

Sweeney, Mike. 2020. "Nearly Two-Thirds of B2B Companies Outsource Marketing—Here's Why." *RightSource*, May 28. https://www.rightsourcemarketing.com/marketing-strategy/why-two-thirds-b2b-companies-outsource-marketing-2/.

Telser, Lester G. 1964. "Advertising and Competition." *Journal of Political Economy* 72 (6): 537–62.

Thumin, F.J. 1962. "Identification of Cola Beverages." *Journal of Applied Psychology* 46 (5): 358–60.

Tuchman, Anna E. (2019). "Advertising and Demand for Addictive Goods: The Effects of E-Cigarette Advertising." *Marketing Science* 38 (6): 913–1084.

Ursu, Raluca M. 2018. "The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions." *Marketing Science* 37 (4): 530–52.

Visser, Jody, Alannah Sheerin, and Field. 2018. "Is It Time to Bring More of Your Marketing In-House?" *BCG*, July 31. https://www.bcg.com/publications/2018/time-to-bring-more-your-marketing-in-house.

Zoltners, Andris A., P.K. Sinha, and Sally E. Lorimer. 2013. "Are You Paying Enough Attention to Your Sales Force?" *Harvard Business Review*, April 12. https://hbr.org/2013/04/are-you-paying-enough-attention-to.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 75–102

Four Facts about Human Capital

David J. Deming

Human capital theory is the now widely accepted idea that education, training, and other forms of learning are investments that pay off in the future. Like any capital investment, the costs of schooling are paid up front and the benefits are earned later. To be sure, schooling has benefits far beyond its monetary value, but the relationship between schooling quantity (and quality) and future earnings is one of the most robust findings in social science.

The term "human capital" was initially controversial among the pioneers of human capital theory, who wanted to reject explicitly the implication that people should be treated as property, or that workers are assets who in any sense "belong" to the owners of capital (Goldin and Katz 2020). Despite initial discomfort over terminology, the study of human capital has blossomed. This is in part because people all around the world spend much more money and time on education than they did a half-century ago. Between 1950 and 2010, the share of the world adult population with at least some secondary school education increased from 13 percent to 51 percent, and the share with some tertiary education increased nearly sevenfold, from 2.2 percent to 14.6 percent (Lee and Lee 2016). In the United States, education spending increased from 3.1 percent of GDP in 1950 to 7.1 percent in 2018, with most of the increase coming from the public sector (Digest of Education Statistics 2019, Table 106.10). This pattern generally holds for other countries around

■ *David J. Deming is the Isabelle and Scott Black Professor of Political Economy at the Harvard Kennedy School, Professor of Education and Economics at the Harvard Graduate School of Education, and Research Associate, National Bureau of Economic Research, all in Cambridge, Massachusetts. His email address is david_deming@harvard.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.75.

the world, with faster increases in public spending on education in developing countries (Roser and Ortiz-Ospina 2016).

Research interest in human capital within the economics profession has grown explosively in the last few decades. I conducted a text and title search on EconLit for the phrases "human capital," "education," and "skill." At the time of the seminal work of Becker (1962), about 2.5 percent of articles included at least one of those phrases. This share didn't change much until about 1990, but then it started rising, reaching about 15 percent of all articles since 2015.

This paper synthesizes what we have learned about human capital since Becker (1962) into four stylized facts.[1] First, human capital explains a substantial share of the variation in labor earnings within and across countries. Second, human capital investments have high economic returns throughout childhood and young adulthood. Third, the technology for producing foundational skills such as numeracy and literacy is well understood, and resources are the main constraint. Fourth, higher-order skills such as problem-solving and teamwork are increasingly economically valuable, and the technology for producing them is not well understood.

We have made substantial progress toward validating the empirical predictions of human capital theory. We know how to improve foundational skills like numeracy and literacy, and we know that investment in these skills pays off in adulthood. However, we have made much less progress on understanding the human capital production function itself. While we know that higher-order skills "matter" and are an important element of human capital, we do not know why.

Fact 1: Human Capital Explains a Substantial Share of the Variation in Labor Earnings within and across Countries.

The "Mincer equation," as it is colloquially known, is an important building block of human capital theory. Mincer (1974) starts with a formal model where identical agents make forward-looking investments in human capital to maximize the present value of future earnings and derives this relationship:

$$\ln y_i = \alpha_i + \beta S_i + \gamma X_i + \delta X_i^2 + \varepsilon_i.$$

The Mincer equation models log annual earnings (or sometimes hourly wages) y_i as an additive function that is linear in years of schooling S_i and quadratic in years of experience X_i. Although subsequent work has proposed adding higher-order terms in experience and nonlinearities in education, the Mincer equation has mostly withstood the test of time (Lemieux 2006).

[1] The structure of the article models Nicholas Kaldor's (1961) six "stylized" facts about economic growth as well as the six "new Kaldor facts" discussed in Jones and Romer (2010). Both articles summarized the state of knowledge about economic growth and successfully framed the research agenda going forward. That is also my goal for research on human capital.

The Mincer model's simple functional form spawned a large literature of different approaches to estimating β, the economic return to an additional year of schooling. Across many different countries and settings, estimates of β yield a coefficient of around 0.1, which implies that another year of schooling increases earnings by 10 percent (Gunderson and Oreopoulos 2020; Patrinos and Psacharopoulos 2020).

One immediately apparent issue is the potential endogeneity of schooling. Rational agents will invest more in schooling when they expect to receive higher returns, and thus a naïve comparison of earnings between individuals with different amounts of completed education will suffer from "ability bias" as noted by Griliches (1977), Card (1999), and many others.

One solution is to find an instrumental variable that affects schooling but is unrelated to ability or other determinants of earnings. The search for such instrumental variables has been a central focus for labor economists over the past few decades. Possibilities include distance to the nearest college (Card 1995), compulsory schooling laws that vary across countries and states and over time (Angrist and Krueger 1991), the timing of school construction (Duflo 2001), and the expansion of funding for primary schools (Khanna 2021).

An alternative "regression discontinuity" approach uses discontinuous changes in the probability of admission around grade or test thresholds to estimate returns to education. Zimmerman (2014) finds that students with a grade point average just high enough to be admitted to Florida International University have 22 percent higher earnings a decade after they apply. This translates into an 11 percent return for a year of education if there is no return to community college, or an 18 percent return if the plausible alternative for many of these students—a year of community college—is worth the same as a year at Florida International University.[2] Several other studies find positive earnings impacts of admission to high schools and colleges, with some emphasizing the quantity and others the quality of education.[3]

The bottom line is that naïve cross-sectional comparisons and studies with strong quasi-experimental research designs yield very similar estimates of the economic return to education. Overall, studies that identify returns to education using instrumental variables, regression discontinuity, and other quasi-experimental approaches yield estimates of an additional year of education ranging between 6 and 18 percent, with a median in the 10–12 percent range. This is slightly higher than the 10 percent return from a "naïve" Mincer model, most likely because of

[2]The admissions standards at Florida International University were more generous than any other four-year public university in Florida, and so students who did not meet that admissions threshold mostly attended community colleges or did not go to college at all. Barrow and Malamud (2015) calculate that the return to a year of college would be 18 percent if the earnings difference around the threshold in Zimmerman (2014) was due entirely to the difference in average years enrolled.

[3]For example, Hoekstra (2009) finds that white men have 20 percent higher earnings when they barely meet an admissions threshold at a state flagship university. Canaan and Mouganie (2018) find that students who marginally pass a French high school exit exam enroll in higher-quality colleges and earn 12.5 percent more, despite no increase in the quantity of education.

some combination of measurement error and higher returns for marginal students (Card 1999). Across all OECD countries, the median earnings premium for a four-year college/tertiary education is 52 percent, or roughly 13 percent per year of education.[4]

Card (1999) finds that a standard Mincer model with a linear schooling term explains between 20 and 35 percent of the variation in labor earnings using the Current Population Survey, a cross-sectional survey of US workers. However, it is not possible with this data to follow workers over the life course or to account for possible "ability bias" in the return to schooling.

Table 1 uses data from the 1979 National Longitudinal Survey of Youth (NLSY79), which tracks a cohort of youth ages 14 to 22 in 1979 as they progress through the labor market. To estimate returns to education over the life course, I compute the average inflation-adjusted hourly wage for individuals between the ages of 25 and 54 over multiple observations, and then regress log average hourly wages on years of education, race and gender indicators, and cognitive ability as measured by adolescent scores on the Armed Forces Qualifying Test (AFQT).

Column 1 shows that the average return to a year of education over an individual's prime working years is 10.9 percent. The R^2 of this regression is 30 percent. Controlling for AFQT scores to account for "ability bias" reduces the coefficient on years of education to 7.2 percent and increases the R^2 of the regression to 35 percent. Column 3 shows the average return for different levels of educational attainment, with less than high school as the left-out category. High school graduates and four-year college graduates earn an average of 13 percent and 48 percent higher wages than those with less than a high school education, respectively. These results are similar in magnitude to the quasi-experimental studies discussed above and to naïve cross-sectional estimates from other data sources. Basic measures of human capital such as education and cognitive ability can explain at least one-third of the variation in wages in a recent cohort of US workers.[5]

However, one-third is probably a lower bound for the impact of human capital on the variance in earnings, for three reasons. First, the calculation here does not include variation in education quality between workers with the same level of attainment. Quality adjustment is particularly important, because nearly all expansion of US postsecondary education over the last few decades has occurred within less-selective institutions. Carneiro and Lee (2011) estimate that the college premium would have grown an additional 30 percent between 1960 and 2000 if pre-college education quality were held constant for the marginal college graduate.

Second, several studies find a larger role for human capital when it is measured in a way that includes education, but also other attributes. For example, Smith et al. (2019) study the impact of owner death or retirement on private

[4] Based on OECD.Stat data at https://stats.oecd.org/Index.aspx?DataSetCode=EAG_EARNINGS.

[5] Hoffmann, Lee, and Lemieux (2020) estimate that education is responsible for more than half of the growth in earnings inequality in the United States since the 1970s, and nearly 75 percent of the growth in inequality since the late 1980s.

Table 1
Returns to Education in the NLSY79

	Average Log Hourly Wages		
	(1)	(2)	(3)
Years of education	0.109 [0.002]	0.072 [0.002]	
AFQT (standardized)		0.161 [0.006]	0.154 [0.006]
High school graduate			0.127 [0.013]
Some college			0.247 [0.016]
Bachelor's degree			0.479 [0.019]
Graduate degree			0.535 [0.021]
R^2	0.296	0.346	0.348
Sample size	10,876	10,876	10,876

Note: Estimates are from a regression of inflation-adjusted average log hourly wages, measured between the ages of 25 and 54 using repeated observations of individuals in panel data from the 1979 National Longitudinal Survey of Youth (NLSY79). In columns 1 and 2, years of education is a continuous measure that is bounded below at 11 and above at 20. Column 3 shows results by level of attainment, where less than high school is the left-out category. Demographics are indicators for race and gender. AFQT is the Armed Forces Qualifying Test, a measure of aptitude administered prior to labor market entry and standardized to have mean zero and standard deviation one. The average wage of respondents is $18.95 in 2016 dollars.

pass-through businesses and find that 75 percent of profits are attributable to the owner's human capital, rather than physical or financial assets. Card et al. (2018) and Song et al. (2019) decompose the variance of earnings in matched employer-employee data and find that "worker effects" account for 40 percent of the variance in earnings in West Germany and 50 percent in the United States, respectively. Because worker effects are invariant to firm pay premia and occupational shifts by construction, we can reasonably consider them an estimate of workers' human capital.

Third, there is the possibility of human capital externalities, where one person's education increases the earnings of others around them. The literature on human capital externalities is not settled, with some studies finding little or no evidence and others finding relatively large agglomeration effects of working in geographic areas or firms with higher levels of human capital (Acemoglu and Angrist 2001; Moretti 2004; Ciccone and Peri 2006; Gennaioli et al. 2012). Externalities, if they exist, would only increase the importance of human capital for explaining variation in labor earnings.

Some authors argue that schooling simply reflects higher human capital, rather than causing it (Caplan 2019). The signaling model of Spence (1974) suggests that individuals invest in education because of the information value it sends to

employers about their productivity. It is difficult to disentangle human capital and signaling empirically, and both explanations surely contribute somewhat to explaining returns to education. However, I think the contribution of signaling is probably small, for two reasons.

First, many studies find positive returns to education even when no degree or credential is earned. This is important because signaling theory requires employers to observe the signal, and most people don't report years of education on a resume. For example, studies of compulsory schooling compare groups of students who all seek to drop out as soon as they can, but some are required to stay in school longer based on when they were born during a calendar year. Many of the youth staying in school for an extra year do not end up obtaining a high school degree at all—they drop out in 11th grade rather than 10th grade. Nonetheless, such studies show that additional education leads to gains in earnings. A school construction program in Indonesia studied by Duflo (2001) mostly worked by increasing primary school enrollment, not receipt of degrees—but still led to later gains in wages. Aryal, Bhuller, and Lange (2022) cleverly exploit the differential observability of compulsory schooling laws across regions in Norway to separate returns to human capital from signaling, and find that human capital accounts for 70 percent of the private return to secondary school education.

Various studies find large labor market returns to increased coursework requirements and specific skills and knowledge learned in high school or college, even if they do not lead to more degrees being earned (for example, Arteaga 2018). Moreover, some fields such as engineering, law, and medicine impart concrete skills and specialized knowledge differences that self-evidently reflect human capital accumulation. No one was born knowing how to be a heart surgeon.

Second, empirical support for signaling theory is scant. Clark and Martorell (2014) find no difference in earnings between high school students who barely pass or fail an exit exam, implying that there is no signaling value of a high school diploma. Some studies do find that the return to education decreases over time as employers learn workers' true ability, which is a testable implication of the signaling model (Altonji and Pierret 2001; Lange 2007). Yet a similar test of the employer learning model in a more recent cohort finds that the return to education does not diminish with experience (Castex and Dechter 2014).

The evidence described above suggests that human capital explains at least one-third of the variation in labor earnings within the United States. How much of the *cross-country* variation in earnings can be explained by human capital?

Following Solow (1956) and Hall and Jones (1999), a standard approach here is to look at an aggregate production function for the economy, where total output is expressed in terms of inputs of quality-adjusted human capital, physical capital, and technology.[6] While data on output, education, physical capital, and the labor

[6] Hall and Jones (1999) consider an aggregate production function for the economy written in terms of log output per worker:

$$\ln\left(\frac{Y_c}{L_c}\right) = \frac{\alpha}{1-\alpha}\ln\left(\frac{K_c}{Y_c}\right) + \ln\left(\frac{H_c}{L_c}\right) + \ln\left(\frac{A_c}{L_c}\right)$$

force are widely available, data on technology is not, and so the measurement of technology—often called total factor productivity—shows up in cross-country studies as the "Solow residual" that is not explained by other measured variables. Mankiw, Romer, and Weil (1992) show that countries with higher rates of human capital grow faster, and that human capital is positively related to GDP growth over a 25-year period.

However, just as in the Mincer model, cross-country differences in schooling are probably endogenous: that is, countries with better technology will benefit more from investments in human capital and thus tend to make such investments more often, and so causality cannot be inferred from the basic relationships.[7] The solution in the individual case involves seeking out methodologies or experiments that change the level of schooling, holding other factors constant. For cross-country differences, the ideal experiment would vary a country's human capital stock or its total factor productivity, holding the other factors constant.

Hendricks and Schoellman (2018) approximate this experiment by studying the wage gains from migration. If skills travel with individuals when they migrate, relative wages across countries with different technologies and institutions can inform us about the contribution of human capital to cross-country income differences. Hendricks and Schoellman (2018) measure pre- and post-migration wages of US migrants using the New Immigrant Survey. They find that migrants to the United States from low-income countries experience wage gains equal to 38 percent of the total GDP-per-worker gap in each source country. Intuitively, these migrants are experiencing a change in total factor productivity and institutions while their human capital is held constant. If the wage gains from this change are equal to 38 percent of the cross-country difference in GDP-per-worker, the remaining 62 percent is explained by human capital.

Their approach has two potential sources of bias. First, human capital may not transfer fully across countries. However, when they apply their method to immigrants who come to the United States on employment visas, have job offers in hand, and work in the same occupation, they show that human capital still accounts for at least 50 percent of cross-country income differences in these cases. Second, immigrants may be self-selected in the sense that those with an expectation of larger earnings gains may be more likely to migrate. However, selection on gains would bias cross-country earnings differences upward, leading them to understate the importance of human capital. In a follow-up paper, Hendricks, Herrington, and Schoellman (2021) use the wage gains from migration to calibrate models of development

where Y_c represents total output in country c, K is capital and α is the capital share, H is quality-adjusted labor, and A is a term representing the state of technology, often called total factor productivity (TFP). This equation can be estimated using cross-country data on incomes and factor shares, with human capital per worker $\frac{H_c}{L_c}$ measured using years of schooling or other data on educational attainment.

[7] Development accounting estimates of the importance of human capital for economic growth depend greatly on measurement and on the assumed structure of the aggregate production function. Rossi (2018) and Hendricks and Schoellman (forthcoming) are excellent reviews of the literature.

accounting under different assumptions, and estimate that human capital explains between 50 and 75 percent of cross-country income differences.

Overall, the best evidence suggests that human capital accounts for at least one-third of the variation in labor earnings within countries, and at least one-half of the variation in earnings per worker across countries.

Fact 2: Human Capital Investments Have High Economic Returns Throughout Childhood and Young Adulthood.

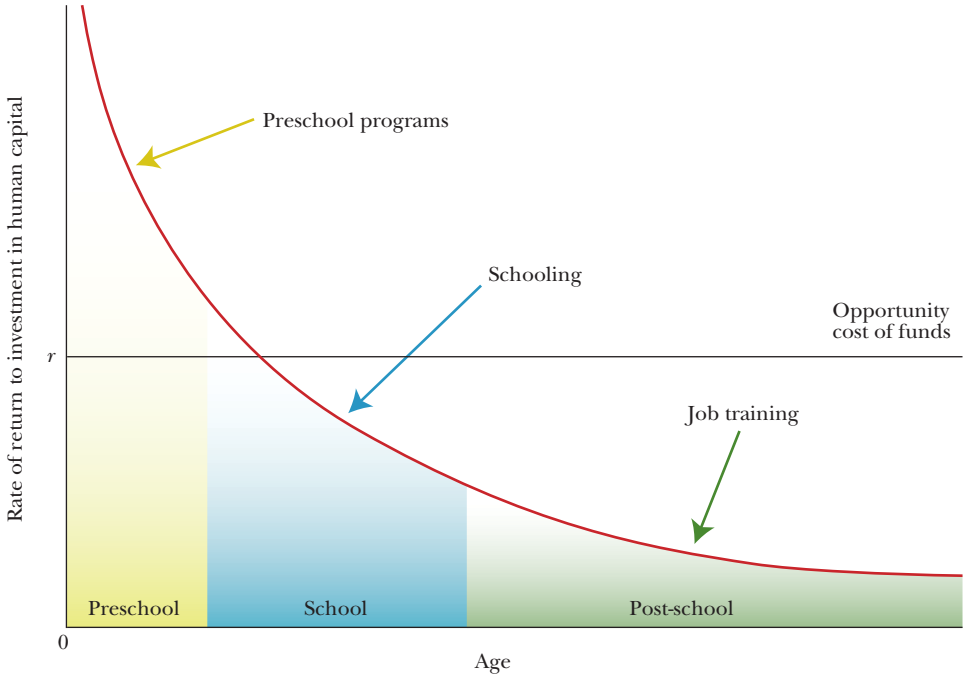
The last few decades have seen increased public support for early childhood investment in the United States and around the world. In the United States, the share of four year-olds enrolled in state-run preschool increased from 15 percent in 2003 to 34 percent in 2019 (National Institute for Early Education Research 2021). Between 1973 and 2014, the number of children in the world enrolled in pre-primary education increased from 43.6 million to 155.1 million (Roser and Ortiz-Ospina 2017). Some of the motivation for this increase is the belief that the payoff to early-life investment is especially high. However, while a body of empirical evidence confirms the high returns to early-life investment, evidence confirms similarly high returns to an array of young adulthood investments as well.

In a series of papers, James Heckman and his co-authors have argued that the economic return to human capital investment diminishes as children age. Figure 1 reproduces the "Heckman curve," a key illustration of the concept of diminishing returns on skill investment (Heckman 2006). The figure shows a declining rate of return, with a horizontal "break even" line for public investment in human capital formation that intersects somewhere during school-age.

Cunha and Heckman (2007) formalize these ideas with a model of life-cycle skill formation. Agents are born with human capital (which could reflect genes, parental education, income, and other fixed factors) and an initial endowment of skills that can expand over time. They consider a general "technology of skill formation" where early investments can matter more than late investments, and where it is not always possible to remediate early skill deficits completely. A key idea from their model is "self-productivity," which is captured by the memorable phrase "skills beget skills." As an intuitive example, self-productivity matters for cumulative learning processes such as mathematics, where concepts build upon one another. More broadly, early childhood investments can raise the level of human capital in a way that increases the productivity of later childhood investments.

Another key idea in the Cunha-Heckman model is "dynamic complementarity." Imagine that there is a fixed budget of skill investment dollars available to be spent on each child. Dynamic complementarity suggests that a balanced investment portfolio yields higher returns than spending lots of money later on and very little early in life, for example. The combination of self-productivity and dynamic

Figure 1
Rates of Return to Human Capital Investment



Source: Heckman (2006, Figure 2).

complementarity implies that later investments are not very productive and that they cannot easily remediate early skill deficits.[8] This model offers a theoretical rationalization of the Heckman curve.

There is strong evidence supporting the value of skill investments in early childhood. Perhaps best-known are two randomized evaluations of preschool interventions from the 1960s: the High/Scope Perry Preschool Project and the Carolina Abecedarian Project. These studies are from decades ago, involving small-scale and intensive interventions for highly disadvantaged families, and thus their results may not generalize to larger and more recent programs. However, several studies of more recent preschool interventions also find substantial impacts—although the evidence also provides a puzzle. Pre-K programs often provide only a short-term boost to test scores that fades out in a few years. Yet they have longer-run impacts on

[8] Cunha and Heckman (2007) propose a two-period constant elasticity of substitution production function for adult skills $A = h\left[\gamma I_1^{\phi} + (1-\gamma) I_2^{\phi}\right]^{1/\phi}$ where $0 \leq \gamma \leq 1$ is a share parameter and $\frac{1}{(1-\phi)}$ is the elasticity of substitution with $\phi \leq 1$. The importance of "self-productivity" is increasing in γ because of the higher relative weight on early life investments. "Dynamic complementarity" is decreasing in ϕ, for example as $\phi \to -\infty$ the production function converges to the perfect complements case, $h[min(I_1, I_2)]$.

important life outcomes such as high school graduation and college attendance, as well as non-educational metrics like reductions in crime and teen pregnancy and improved health later in life (for example, Ludwig and Miller 2007; Deming 2009; Gray-Lobe, Pathak, and Walters 2021).

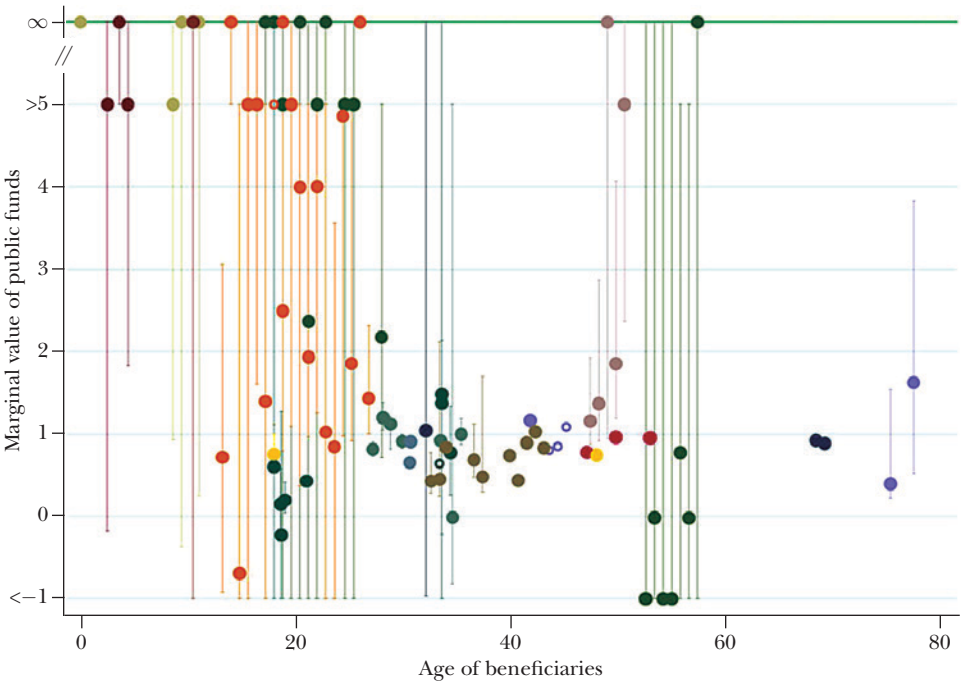
In addition, there are many other methods of early childhood investment: prenatal care, early child health care, food and nutrition support, home visits to encourage practices like breastfeeding and smoking cessation, and others. In recent essays in this journal, Aizer, Hoynes, and Lleras-Muney (2022) describe the evidence of long-term benefits from policy interventions affecting low-income children like Medicaid and food stamps, while Wüst (2022) presents the evidence from the Nordic countries about the benefits of universal provision of early childhood investments in pre-natal care, health care at time of birth, and early childhood health care. Again, any benefit-cost analysis of these programs needs to take a long-term view, because many of these benefits only become apparent later in life.

The Heckman curve has practical implications for policymaking. Heckman (2006) writes that "early interventions targeted toward disadvantaged children have much higher returns than later interventions such as reduced pupil-teacher ratios, public job training, convict rehabilitation programs, tuition subsidies, or expenditures on police."

But of course, evidence of high returns from early childhood interventions does not imply lower returns for other interventions. Indeed, a body of evidence suggests that human capital investments have high returns through childhood and young adulthood. Hendren and Sprung-Keyser (2020) summarize the findings of 133 experimental and quasi-experimental policy interventions in the United States—interventions affecting a wide variety of age groups—using a unified welfare analysis framework. They calculate the "marginal value of public funds" for each of these studies as the ratio of recipients' willingness to pay by the net cost to the government. A marginal value of public funds greater than 1 translates into a social welfare improvement over a non-distortionary cash transfer, and an infinite marginal value of public funds means that the program is a Pareto improvement that "pays for itself" due to the positive fiscal externality created when earnings increases are large enough to pay back program costs through increased tax revenue. For an introduction to the "marginal value of public funds" framework in this journal, see Finkelstein and Hendren (2020).

Figure 2—which is reproduced from Hendren and Sprung-Keyser (2020)—plots marginal value of public funds estimates from all 133 policy interventions, sorted by the age of beneficiaries. Perry Preschool and Carolina Abecedarian are included here and have very high marginal value of public funds. But so do other policies aimed at other age groups, such as increased K–12 school spending in the 1970s and 1980s, financial aid for low-income college students, and sectoral employment programs for young adults. While individual estimates are noisy, they pool studies by category and find that child education, child health, and college policies all "pay for themselves" on average through increased future tax receipts. This pattern of results by age group is not an artifact of their welfare analysis framework,

Figure 2
All Marginal Value of Public Funds Estimates with Confidence Intervals



Source: Hendren and Sprung-Keyser (2020, Figure IV.A).
Note: This figure shows 133 estimates of the Marginal Value of Public Funds (MVPF) for experimental and quasi-experimental interventions, grouped by the age of beneficiaries. An infinite MVPF indicates policies that "pay for themselves" through fiscal externalities such as increased future tax receipts.

because many of the studies they cite also have high returns when measured with more conventional approaches such as benefit-cost ratios or internal rates of return (Rea and Burton 2020).

In sum, the evidence suggests that human capital investments are, at least in rough terms, *equally productive* between the ages of 0 and 25. The key distinction is not age per se, but rather a focus on human capital. Skill investments improve outcomes for adult recipients, including higher income but also improvements in health and other benefits. However, higher income later in life benefits society as well as participants themselves, because the resulting increase in tax revenue lowers the long-run fiscal cost of a program. In contrast, programs for adults such as housing vouchers or disability insurance tend to reduce labor earnings, which pushes the marginal value of public funds below 1. To be clear, this does not mean that the policies are a bad idea, just that supporting such policies requires placing a higher welfare weight, in a given year, on beneficiaries than on the average taxpayer.

What are some reasons why skill investments would have similar returns throughout the life course? First, perhaps even young children are already on the flat of the Heckman curve. Exposure to disease, pollution, and other adverse events have temporary impacts on adults and young children but long-lasting and permanent impacts on fetuses (for example, Almond, Currie, and Duque 2018). Perhaps sensitivity to early investments might be most important before a child is born, with the difference between early and late childhood being less important. Second, human capital investments during school-age may be more productive on the margin because schools are an efficient delivery mode for interventions.[9] High fixed costs require interventions to be very intensive, whereas the marginal dollar can be spent more on increasing dosage when fixed costs are low. School spending and financial aid exploit the fact that the fixed costs have already been paid.

Third, while the Heckman curve arises from an economic theory about the technological possibilities for human capital investment, the real world is much messier. Due to lack of opportunity and various market failures such as credit constraints and imperfect information, few people reach their full potential. If almost everyone is inside their own skill frontier, the Heckman curve may not apply in practice even if it exists in principle.

Fact 3: The Technology for Producing Foundational Skills Such as Numeracy and Literacy Is Well Understood, and Resources Are the Main Constraint.

Human capital investments can be productive at many stages of life. Yet not all interventions increase human capital, and what works in one setting may not work in others. *How* should we invest in human capital?

The available options can be divided into three main categories. First, we can allow schools to improve input quality by relaxing their financial constraints. With extra money, schools might buy smaller classes, higher-quality teachers, additional tutors, or other inputs. Second, we can change the investment decisions of individuals, families, or schools by lowering the relative price (perhaps to zero) of specific inputs like tutoring or technology, or by increasing incentives for specific inputs like coming to class or reading books. Third, we can encourage students, teachers and other actors to expend more effort toward human capital investment through performance incentives.

There has been an ongoing controversy over which of these approaches holds greatest promise. In a highly influential review titled "The Failure of Input-Based Schooling Policies," Hanushek (2003) argues that investments such as lowering class size or increasing teacher pay do not work because schools do not use resources efficiently. He advocates instead for greater performance incentives to increase

[9] I thank Todd Rogers for a helpful conversation that crystallized this point.

teacher and school effort. However, a wave of recent research reviewed in Jackson (2020) has concluded that additional resources do improve education outcomes. The discrepancy between these two findings arises partly from timing, but also from a debate about research design and methodology. Hanushek's argument was based on time-series and cross-sectional differences in education spending, and how it was difficult to find positive correlations in this data between spending and educational outcomes. The newer research is based on quasi-experimental evidence from school finance reforms.

As one example, Jackson, Johnson, and Persico (2016) use court-ordered changes in state K–12 funding formulas to predict a local district's reform-induced expected change in per-pupil spending on class size, instructional time, and teacher quality, and then study the impact of these funding changes on student outcomes. They find that a 10 percent increase in per-pupil spending over 12 years of public schooling increases educational attainment by 0.3 years and adult wages by 7 percent. Jackson and Mackevicius (2021) conduct a meta-analysis of 31 quasi-experimental studies relating US public K–12 school spending to student outcomes. They estimate that a $1,000 increase in per-pupil spending over four years increases test scores by 0.035 standard deviations and increases college-going by 2.6 percentage points.

Additional school spending also boosts human capital in developing countries. Duflo (2001) finds that a large school construction program in Indonesia increased educational attainment and earnings. Khanna (2021) shows that school districts in India that received extra resources because of a funding formula discontinuity built more schools, hired more teachers, and improved existing facilities. Students in these districts completed about 0.7 more years of schooling and earned between 11 and 14 percent more as adults. However, other studies found no effect of increased school spending: de Ree et al. (2018) find no impact of a large increase in teacher salaries in India, and Mbiti et al. (2019) find no impact of unconditional grants to schools in Indonesia.

One partial explanation of these varying results is that school finance reforms operate as "helicopter drops" of additional resources. While there is some heterogeneity in how the money is targeted or the characteristics of the student population, schools mostly seem to use extra resources to do more of what they were already doing. In the United States, at least, "more of the same" can be a good investment on the margin. Hendren and Sprung-Keyser (2020) calculate an infinite marginal value of public spending for the Jackson, Johnson, and Persico (2016) results, suggesting that increased school spending caused by court decisions in the 1970s and 1980s "paid for itself" through increased future tax revenues. However, unconditional resource increases have often worked less well in developing countries: for example, the World Bank (2018) began its *World Development Report* with the sentence: "Schooling is not the same as learning." The report discusses a common pattern across many countries, where resources devoted to getting children into classrooms have not been followed by a commensurate increase in academic achievement.

Which specific input investments reliably increase human capital? In the context of developed economies, Fryer (2017) reviews nearly 200 randomized educational interventions and finds wide and sometimes unpredictable variation in "what works." Experiments that lower poverty, change parenting practices, or alter the home environment have no average impact on academic outcomes. Early childhood and school-based interventions are effective on average, but with substantial heterogeneity. High-dosage tutoring sometimes increases math and reading achievement, but low-dosage tutoring does not (for example, Banerjee et al. 2007; Fryer and Howard-Noveck 2020). Teacher training and professional development sometimes increases achievement and sometimes does not (for example, Borman et al. 2007; Loyalka et al. 2019). Providing computers at home or in school generally has no impact on measured human capital (for example, Malamud and Pop-Eleches 2011; Cristia et al. 2017). Computer-assisted learning software has mixed impacts on achievement, with some evidence that technology-assisted personalization increases student achievement (Bulman and Fairlie 2016; Muralidharan, Singh, and Ganimian 2019).

The overall picture is that some specific input investments work very well, but many do not, and it is often hard to predict ahead of time. A Coalition for Evidence-Based Policy (2013) summary of randomized evidence on targeted school input interventions found that 11 out of 90 interventions produced positive and statistically significant impacts on achievement. Moreover, interventions that are effective in one context may not scale up or generalize to other settings. Kerwin and Thornton (2021) show that while the full-service version of a literacy program in Uganda boosted reading and writing skills, a lower-cost version implemented in the same context resulted in *lower* test scores. Beg et al. (2019) find that an educational technology intervention in a group of middle schools in Pakistan worked when teachers were trained ahead of time on how to use it in the classroom, but *harmed* learning when it was delivered directly to students.

Finally, the evidence on incentives is mixed. In experiments that included 250 urban schools in five US cities, Allan and Fryer (2011) show that paying students for performance directly rarely works. The largest randomized studies of teacher incentives in the United States find no impact on student achievement or other outcomes (Springer et al. 2012; Fryer 2013). Teacher incentives sometimes have positive impacts in developing countries, with larger impacts on tested versus non-tested subjects (for example, Muralidharan and Sundararaman 2011; Filmer, Habyarimana, and Sabarwal 2020).

A large body of research studies school incentives in the form of test-based accountability, which threatens low-performing schools with sanctions such as failing grades, dismissal of teachers and principals, and school closure.[10] This form

[10] Another way to increase schools' effort is through competition between traditional public schools and charter and/or private schools. Figlio, Hart, and Karbownik (2020) find that the introduction of private school vouchers in Florida modestly increased achievement among students attending nearby public schools due to competitive pressure. However, in a large market-level experiment in India, Muralidharan

of accountability pressure leads to large gains on high-stakes tests, modest and inconsistent gains on low-stakes tests, and a variety of harmful strategic responses (for example, Jacob and Levitt 2003; Neal and Schanzenbach 2010; Dee and Jacob 2011). In terms of long-run impacts, Deming et al. (2016) find that accountability pressure in Texas increased college attendance and earnings for students in the lowest-performing schools but had negative long-run impacts for low-scoring students in higher-rated schools due to strategic classification of students as eligible for special education.

When incentives work, they often do so by diverting effort toward the narrow goal of meeting a performance target in ways that can create harmful side effects. This "multi-tasking" problem is well-known in the economics literature (Holmstrom and Milgrom 1991). Incentives increase pressure on students, teachers, and schools to meet short-run targets, when the actual goal is long-run human capital development. In some form, this tradeoff may be unavoidable. Even though education is a long-run process, incentives usually work better for immediate and easy-to-verify metrics like attendance, enrollment, reading books, and completing quizzes.

Some promising evidence suggests that resources and incentives can work well together. In a large-scale experiment in Tanzania, Mbiti et al. (2019) find that teacher incentives and unconditional cash grants to schools have little impact individually, but large impacts on achievement when implemented together. Andrabi, Daniels, and Das (2020) find that unconditional cash grants to private schools in Pakistan increased test scores, but only in villages where the grants were given to all schools rather than only one.

The combination of resources and incentives also works for some high-performing US charter schools that follow a "no excuses" approach, with an emphasis on rules of comportment, longer school days, and extra instructional time. These schools are publicly funded but receive significant additional private funding. Charter schools face higher levels of external accountability because they can be shut down more easily.

I interpret this array of evidence as follows. First, at least in the United States, increased school spending is productive on the margin. Increasing school spending from current levels would produce substantially more human capital and may even pay for itself in the long run. The technology for producing basic math and literacy skills in school-aged children is fairly well-understood. Smaller class sizes, better school facilities, and more instructional time all have reliable impacts on the development of foundational academic skills. The inputs with the best track record of effectiveness—high-dosage tutoring, extra instructional time, personalization, and teaching to the right level—mostly deliver to students "more of the same," rather than reinventing the learning process.

Second, while sharpening incentives works in some contexts, achievement gains are often short-lived and there is not much evidence of long-run benefits. An

---

and Sundararaman (2015) find no impact of private school competition on the achievement of public school students.

important caveat is that resources plus incentives appear to be more effective than resources alone.

Third, simply giving schools money—and allowing them to spend it flexibly—may be a more reliable way to increase human capital than pinning our collective hopes on any particular "silver bullet" approach that all schools would be required to follow. This makes education experts queasy, and rightfully so. In a perfect world, increases in resources are combined with transparency and accountability for results. Yet the evidence suggests that "helicopter drops" of money are spent well enough to be worth the investment, at least in developed countries and in schools with strong internal accountability.

However, just because school spending is economically productive on the margin does not mean that the money is spent optimally. It can be simultaneously true that school spending is productive and that much of it is wasted. We can probably always do better, and so innovation and experimentation are critical for increasing the productivity of human capital investments.

Fact 4: Higher-Order Skills Such as Problem-solving and Teamwork Are Increasingly Economically Valuable, and the Technology for Producing Them Is Not Well Understood.

Schools have a long and successful track record of teaching children how to read, write, and do arithmetic. But a good school does much more. The long-run impacts of educational interventions are often much larger than what would be predicted by achievement gains alone. A growing body of work emphasizes the importance of "non-cognitive" or "soft" skills like patience, self-control, conscientiousness, teamwork, and critical thinking. While such skills are clearly important, the very terms "soft" and "non-cognitive" reveal our lack of understanding about what these skills are and how to measure or develop them.

In my view, the appropriate term for capacities like problem-solving, critical thinking, and teamwork is *higher-order skills*, following Bloom's (1956) taxonomy of educational objectives. Bloom's taxonomy establishes a hierarchy with factual knowledge as the base of the pyramid, followed by pattern recognition and classification, on up to higher-order objectives such as application to new situations, experimentation and making connection to new ideas, evaluation and decision-making, and design and creation of new concepts. Tests like the SAT or the Armed Forces Qualification Test (AFQT) focus on the bottom two layers of the pyramid: recalling, explaining, and understanding ideas and concepts. As discussed above, we know a great deal about how to build these foundational skills: indeed, as the pyramid structure of the taxonomy implies, they are a precondition for developing *higher-order skills* such as applying conceptual knowledge to solve new problems and evaluating evidence from multiple sources to improve decision-making.

Despite our lack of understanding of how higher-order skills are measured and developed, a variety of studies have found ways to use the existing evidence to

demonstrate their importance for life success. For example, Jackson et al. (2020) use survey evidence from ninth-grade students in Chicago public schools and find that schools with high "value-added" in promoting hard work and social well-being increase students' high school graduation and college attendance, even after accounting for their impact on academic achievement. Using data from the population of Swedish military enlistees, Lindqvist and Vestman (2011) estimate high labor market returns to both cognitive and non-cognitive skills, where the latter is measured using scores from a personal interview administered by a trained psychologist. Deming (2017) shows that the economic return to social skills in the United States more than doubled for a cohort of youth entering the labor market in the 2000s compared to the 1980s. In that study, discussed further below, I measure social skills by creating an index based on four factors: self-reported sociability; self-reported sociability at age six, as perceived by the adult respondent; number of clubs in which the respondent participated in high school; and participation in high school sports. Edin et al. (2022) find similar returns to social skills in Sweden, using administrative data from the compulsory military draft that required men aged 18 or 19 to be tested on cognitive and non-cognitive skills. Attanasio et al. (2020) find growing inequality in socio-emotional skills across two British cohorts born 30 years apart, using survey tools filled out by mothers (or in some cases teachers) about behaviors of their children. Each of these studies measures "non-cognitive" skills using whatever measures are at hand, rather than relating them conceptually to particular higher-order skills.

Higher-order skills clearly seem important, yet measuring them well is a challenge. The typical approach uses self-reported questionnaires, which are often Likert scale items (1 to 5 or 1 to 7, ranging from "strongly disagree" to "strongly agree") without any cardinal meaning. Their predictive power for different life outcomes varies widely depending on the exact measure, the outcomes used, and the social context.

One problem is that questionnaires can suffer from reference bias, where respondents make relative comparisons to those around them. West et al. (2016) find that students who win a lottery to attend "no excuses" charter schools subsequently score lower on measures of conscientiousness and grit, despite having higher achievement and attending a school with longer hours and more homework, because they are now evaluating themselves in the context of a different set of institutional expectations. Some studies measure non-cognitive skills using behavior such as absences and school suspensions. Yet such behaviors capture not only skills but also social context, including racial discrimination, school context, and other unknown factors.

Conceptual clarity is a second challenge to our understanding of higher-order skills. In a standard human capital framework, more skills are always better. But certain skills may be effective in some contexts and counterproductive in others: as one example, conscientiousness positively predicts educational attainment and earnings, yet disruptive and aggressive behavior (the opposite of conscientiousness) sometimes predicts earnings and entrepreneurial success (Levine and Rubinstein 2017; Papageorge, Ronda, and Zheng 2019).

We need a systematic research program that seeks to understand the economic importance of higher-order skills. This research program would combine careful measurement and development of theory with experimentation and impact analysis using strong research designs. In the rest of this section, I illustrate the value of this approach by attempting to synthesize what we know (and do not know) about interpersonal and intrapersonal skills.

Interpersonal Skills and the Science of Teamwork

A survey of workers, employers, and experts administered by the US Department of Labor found that teamwork is a "very" or "extremely" important job feature in 78 percent of all jobs (O*NET OnLine 2022). A long literature in economics treats teamwork as a tradeoff between the benefits of increased productivity through specialization and the costs of coordination (Becker and Murphy 1992; Garicano and Rossi-Hansberg 2006). In this context, the rise of team production is a response to the increasing complexity of work, and a well-functioning team can exploit comparative advantage between team members to increase productive efficiency.

Deming (2017) shows how the rise of teamwork has increased the value of social skills in the labor market. Between 1980 and 2012, jobs requiring high levels of social interaction grew by nearly 12 percentage points as a share of the US labor force, and the labor market return to social skills more than doubled. Deming explains these empirical results with a model of team production where social skills reduce the coordination costs of "trading tasks" between workers on a team, allowing them to exploit comparative advantage more fully. Several other recent papers show evidence of the economic value of social skills. For example, Hansen et al. (2021) use data on job descriptions for top executives to "show an increasing relevance of social skills in top managerial occupations." Hoffman and Tadelis (2021) look at employee surveys within a single large firm to show that managers with better "people management skills" reduce attrition among those working for them and are compensated better by the firm.

We know that social skills are rewarded in the labor market, and we know that teamwork is increasingly important. But can we draw a direct connection between social skills, teamwork, and increased productivity?

Weidmann and Deming (2021) develop a novel experimental method for identifying individual contributions to group performance. We first measure individuals' productivity on a series of problem-solving tasks, and then randomly assign the same individuals to multiple teams, which perform group analogs of the same tasks.[11] We use the individual scores to generate a performance prediction for each

[11] The "memory test" involving words, images, and stories. In the "optimization test," participants made a series of guesses between 0 and 300, observed how these guesses were translated by a complex unobserved function into final values, and then estimated the highest value for the function. In the "shapes test," participants observed a series of shapes and then predicted what element was missing in the next shape of the series.

team, and then estimate a "team player" effect by combining the residual from the prediction across multiple random assignments of individuals to groups. People who consistently cause their teams to outperform its prediction are team players. We find that individuals have persistent impacts on group performance—in other words, that team-player effects exist.

In addition, these effects found in Weidmann and Deming (2021) are positively correlated with a commonly used and psychometrically validated measure of social intelligence called the Reading the Mind in the Eyes Test (Baron-Cohen et al. 2001). The test presents participants with photos of faces, cropped so that only the eyes are visible. For each set of eyes, participants are asked to choose which of four emotions best describes the person in the image. This test measures the ability of participants to recognize emotions in others and, more broadly, the ability to reason about the mental state of others. Relative to other measures of social intelligence, the main value of the Reading the Mind in the Eyes is that it has right and wrong answers, has relatively high test-retest reliability, and can be quickly and reliably administered (Pinkham et al. 2014). Lab participants were also assessed on a standard measure of IQ and on three dimensions of the well-known "Big 5" personality inventory that are positively associated with group performance in other studies: Conscientiousness, Extraversion, and Agreeableness. None of these measures were correlated with the team player effect—only the Reading the Mind in the Eyes Test. In short, this experiment uses a lab setting to develop a clean test of the underlying economic mechanism relating social skills to team productivity.

The importance of teamwork skills is also corroborated in a variety of field settings. Arcidiacono, Kinsler, and Price (2017) use data from US professional basketball to show how individual performance depends on peer effects. Devereux (2021) looks at data on co-authorship of academic papers for economists within the University of California system; he finds that the importance of an author's value-added as a co-author is more closely tied to salary than the author's own value-added. Bonhomme (2021) develops an econometric framework to estimate the impact of individual workers on team performance that allows for variation in teamwork skills, worker sorting, and complementarity, and estimates the framework using the research output of economists and contributions of inventors to patent quality. However, much more work is needed to understand how social skills matter and under what conditions.

Another largely unexplored frontier is the *development* of social skills. There are a few studies, often in developing economies, looking at how a specific social skills program improved outcomes. With female workers in the garment industry in India, Adhvaryu, Kala, and Nyshadham (forthcoming) find that on-the-job "soft skills"—with a focus on communication, time management, financial literacy, successful task execution, and problem-solving—increases employee productivity by 13.5 percent, with larger impacts when work is more team-intensive and evidence of spillovers to untreated teammates. In a business training program in Togo, social skills training programs improve entrepreneurs' ability to form business connections (Dimitriadis

and Koning 2020). In a three-week business skills program for high school students in Uganda, Chioda et al. (2021) find that teaching soft skills like self-efficacy, persuasion, and negotiation led to greater gains in earnings than a focus on hard skills. In an educational setting in Zambia, Ashraf et al. (2020) find that a training program in negotiation skills for adolescent girls improved educational outcomes. We need more research and varying programs and contexts to build an economic theory of how and why teamwork skills matter.

Intrapersonal Skills and Economic Decision-Making

Good decision-making requires counterfactual reasoning, meaning a consideration of alternative actions and their likely consequences. In terms of Bloom's (1956) taxonomy, this process requires the combination of several higher-order skills such acquiring information, applying information to new situations, testing and evaluating evidence, and making and justifying decisions.

A wide literature has considered various and overlapping aspects of decision-making: patience, self-control, grit to persevere through difficult tasks, habits of acquiring additional information or considering alternative strategies, developing cognitive shortcuts that will make sense in a variety of settings, and others. The economics literature on time and risk preference has mostly focused on broad cultural and familial influences for these decisions, or on behavioral biases and attributes of the choice environment. Yet in my view, decision-making errors and biases can be reinterpreted as arising from deficits in higher-order skills. This in turn suggests the need for more research on the "skill" of decision-making itself.

For example, patience and self-control are linked: patience is a willingness to think long-term, whereas self-control is the ability to overcome the temptations of the present (Fudenberg and Levine 2006). One approach in this literature asks the subject to compare a tradeoff between now and the future to two equivalently spaced times in the future: for example, the offer of $100 today versus $110 tomorrow is accepted much more often than the offer of $100 seven days from now versus $110 eight days from now. Waiting an extra day seems easier in the future than in the present. Self-control is positively related to academic achievement and other life outcomes: as one example, Duckworth et al. (2019) survey a range of evidence that measures of self-control are linked to academic achievement. Related to self-control is "grit," or the ability to persevere through effortful, sometimes unpleasant tasks to achieve a desired long-run outcome.

There is promising evidence that self-control and grit are malleable and can be improved through low-cost, scalable investments. Alan and Ertac (2018) show that a school-based enrichment program for third- and fourth-graders in Turkey that encourages forward-looking behavior and imagining the future increases behavior grades and patience up to three years later. Lührmann, Serra-Garcia, and Winter (2018) show that a financial literacy program for German high school students reduced time inconsistency and lowered discount rates.

What is the best way to improve grit? One view is that grit is developed through the adoption of a "growth mindset" (for example, Yeager and Dweck 2012), which

refers to changes in beliefs about the returns to effort. An alternative view treats grit as *cognitive endurance*, the skill of maintaining focus over time (Brown et al. 2021). Interventions focusing solely on growth mindset have shown mixed results, yet pairing mindset training with structured support for goal-setting and deliberate practice yields promising increases in academic performance in a nationally representative group of US high school students, as well as in 52 state-run elementary schools in Turkey (Yeager et al. 2019; Alan, Boneva, and Ertac 2019). Brown et al. (2021) randomly assign a group of 1,600 Indian primary school students to two types of effortful cognitive activity—one that is clearly academic, and one that is not. *Both* interventions increase the ability to concentrate and both lead to increased academic performance, suggesting that cognitive endurance improves with practice. This reframes "grit" as a skill that can be developed rather than a mindset to be shifted.

Assessing alternative future states of the world is cognitively taxing, which may help explain the correlation between intelligence and patience (for example, Dohmen et al. 2018). One route to better decision-making is to develop cognitive shortcuts and habits of mind that make long-run thinking easier. The success of cognitive behavioral therapy in reducing violence and other negative behaviors offers an intriguing example. Cognitive behavioral therapy focuses on decision-making directly by asking people to slow down and evaluate the consequences of their behavior patterns, and then reprogramming new behaviors through deliberate practice. Heller et al. (2017) find large reductions in violent crime and increases in high school graduation from cognitive behavioral therapy interventions with high-risk young men in Chicago.[12] Blattman, Jamison, and Sheridan (2017) find that a combination of cognitive behavioral therapy and cash grants greatly reduced crime and violence among criminally engaged men in Liberia. A decision-making intervention for young children in Switzerland called Promoting Alternative Thinking Strategies (PATHS) increased high school graduation and college attendance (Sorrenti et al. 2020).

Finally, an emerging body of evidence suggests that strategic sophistication improves decision-making. Fe, Gill, and Prowse (forthcoming) worked with data from the Avon Longitudinal Study of Parents and Children (ALSPAC), which measured the theory-of-mind ability and cognitive ability at age eight of children from the Avon region in the southwest of England, and found that "theory of mind"—the ability to attribute mental states to others—predicts strategic sophistication in children and is positively correlated with adult social skills and educational attainment. Gill and Prowse (2016) find that more intelligent US college students

[12] Heller et al. (2017) give the example of an exercise where students are paired up and one is given a ball. The other is given 30 seconds to try to get the ball from his partner. After 30 seconds of physical struggle, the group leader asks whether anyone decided instead to simply ask for the ball. When they say "no," the leader then asks the ball holder whether they would have given it up, to which the typical answer is "I would have given it; it's just a stupid ball."

converge to Nash equilibrium faster and engage in a more sophisticated level of reasoning in a series of laboratory experiments.

Future studies should seek to develop theory and measurement paradigms that allow for a direct assessment of the skills and knowledge that are required to improve decision-making. One promising approach is to build on the "rational inattention" literature, which identifies the conditions under which decision mistakes are optimal given the costs of paying attention (for example, Sims 2003; Maćkowiak, Matějka, and Wiederholt 2021). Many of the biases and rules-of-thumb phenomena identified by behavioral economics can be rationalized by models of costly information acquisition. Viewed in this light, interventions that build the "skill" of lowering attention costs will manifest as a reduction in decision errors and an increase in patience, grit, and other higher-order skills. When it comes to understanding the role of skills in improving economic decision-making, there are more questions than answers, which suggests many fruitful and exciting avenues for future work.

References

Acemoglu, Daron, and Joshua Angrist. 2000. "How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws." *NBER Macroeconomics Annual* 15: 9–59.

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham. Forthcoming. "Returns to On-the-Job Soft Skills Training." Journal of Political Economy. https://www.dropbox.com/s/9q7tz46g66n7s2j/PACE_apr2021.pdf?dl=0.

Aizer, Anna, Hilary Hoynes, and Adriana Lleras-Muney. 2022. "Children and the US Social Safety Net: Balancing Disincentives for Adults and Benefits for Children." *Journal of Economic Perspectives* 36 (2): 149–74.

Alan, Sule, Teodora Boneva, and Seda Ertac. 2019. "Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit." *Quarterly Journal of Economics* 134 (3): 1121–62.

Alan, Sule, and Seda Ertac. 2018. "Fostering Patience in the Classroom: Results from Randomized Educational Intervention." *Journal of Political Economy* 126 (5): 1865–911.

Allan, Bradley M., and Roland G. Fryer, Jr. 2011. "The Power and Pitfalls of Education Incentives." Hamilton Project Discussion Paper 2011–07.

Almlund, Mathilde, Angela Lee Duckworth, James J. Heckman, and Tim D. Kautz. 2011. "Personality Psychology and Economics." NBER Working Paper 16822.

Almond, Douglas, Janet Currie, and Valentina Duque. 2018. "Childhood Circumstances and Adult Outcomes: Act II." *Journal of Economic Literature* 56 (4): 1360–446.

Altonji, Joseph G., and Charles R. Pierret. 2001. "Employer Learning and Statistical Discrimination." *Quarterly Journal of Economics* 116 (1): 313–50.

Andrabi, Tahir, Benjamin Daniels, and Jishnu Das. 2020. "Human Capital Accumulation and Disasters: Evidence from the Pakistan Earthquake of 2005." Research on Improving Systems of Education Working Paper 20/039.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106 (4): 979–1014.

Arcidiacono, Peter, Josh Kinsler, and Joseph Price. 2017. "Productivity Spillovers in Team Production: Evidence from Professional Basketball." *Journal of Labor Economics* 35 (1): 191–225.

Arteaga, Carolina. 2018. "The Effect of Human Capital on Earnings: Evidence from a Reform at Colombia's Top University." *Journal of Public Economics* 157 (C): 212–25.

Aryal, Gaurab, Manudeep Bhuller, and Fabian Lange. 2022. "Signaling and Employer Learning with Instruments." *American Economic Review* 112 (5): 1669–702.

Ashraf, Nava, Natalie Bau, Corinne Low, and Kathleen McGinn. 2020. "Negotiating a Better Future: How Interpersonal Skills Facilitate Intergenerational Investment." *Quarterly Journal of Economics* 135 (2): 1095–151.

Attanasio, Orazio, Richard Blundell, Gabriella Conti, and Giacomo Mason. 2020. "Inequality in Socio-Emotional Skills: A Cross-Cohort Comparison." *Journal of Public Economics* 191 (November): 104171.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122 (3): 1235–64.

Baron-Cohen, Simon, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. 2001. "The 'Reading the Mind in the Eyes' Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-Functioning Autism." *Journal of Child Psychology and Psychiatry* 42 (2): 241–51.

Barrow, Lisa, and Ofer Malamud. 2015. "Is College a Worthwhile Investment?" *Annual Review of Economics* 7: 519–55.

Becker, Gary S. 1962. "Investment in Human Capital: A Theoretical Analysis." *Journal of Political Economy* 70 (5): 9–49.

Becker, Gary, and Kevin M. Murphy. 1992. "The Division of Labor, Coordination Costs, and Knowledge." *Quarterly Journal of Economics* 107 (4): 1137–60.

Beg, Sabrin A., Adrienne M. Lucas, Waqas Halim, and Umar Saif. 2019. "Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not." NBER Working Paper 25704.

Blattman, Christopher, Julian C. Jamison, and Margaret Sheridan. 2017. "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia." *American Economic Review* 107 (4): 1165–206.

Bloom, Benjamin S., ed. 1956. *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. 2nd ed. London: Addison Wesley Publishing.

Bonhomme, Stéphane. 2021. "Teams: Heterogeneity, Sorting, and Complementarity." Becker Friedman Institute for Economics Working Paper 2021–15.

Borman, Geoffrey D., Robert E. Slavin, Alan C. K. Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers. 2007. "Final Reading Outcomes of the National Randomized Field Trial of Success for All." *American Educational Research Journal* 44 (3): 701–31.

Brown, Christina, Geeta Kingdon, Heather Schofield, and Supreet Kaur. 2021. "Cognitive Endurance as Human Capital." Unpublished.

Bulman, G., and R. W. Fairlie. 2016. "Technology and Education: Computers, Software, and the Internet." In *Handbook of the Economics of Education*, Vol. 5, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 239–80. Amsterdam: Elsevier.

Canaan, Serena, and Pierre Mouganie. 2018. "Returns to Education Quality for Low-Skilled Students: Evidence from a Discontinuity." *Journal of Labor Economics* 36 (2): 395–436.

Caplan, Bryan. 2019. *The Case Against Education*. Princeton: Princeton University Press.

Card, David. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, edited by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky. Toronto: University of Toronto Press.

Card, David. 1999. "The Causal Effect of Education on Earnings." *Handbook of Labor Economics*, Vol. 3, edited by Orley C. Ashenfelter and David Card, 1801–63. Amsterdam: Elsevier.

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.

Carneiro, Pedro, and Sokbae Lee. 2011. "Trends in Quality-Adjusted Skill Premia in the United States, 1960–2000." *American Economic Review* 101 (6): 2309–49.

Castex, Gonzalo, and Evgenia Kogan Dechter. 2014. "The Changing Roles of Education and Ability in Wage Determination." *Journal of Labor Economics* 32 (4): 685–710.

Chioda, Laura, David Contreras-Loya, Paul Gertler, and Dana Carney. 2021. "Making Entrepreneurs: Returns to Training Youth in Hard Versus Soft Business Skills." NBER Working Paper 28845.

Ciccone, Antonio, and Giovanni Peri. 2006. "Identifying Human-Capital Externalities: Theory with Applications." *Review of Economic Studies* 73 (2): 381–412.

Clark, Damon, and Paco Martorell. 2014. "The Signaling Value of a High School Diploma." *Journal of Political Economy* 122 (2): 282–318.

Coalition for Evidence-Based Policy. 2013. "Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive Versus Weak or No Effects." Washington, DC: Coalition for Evidence-Based Policy.

Cohen, Jonathan, Keith Marzilli Ericson, David Laibson, and John Myles White. 2020. "Measuring Time Preferences." *Journal of Economic Literature* 58 (2): 299–347.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín. 2017. "Technology and Child Development: Evidence from the One Laptop per Child Program." *American Economic Journal: Applied Economics* 9 (3): 295–320.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97 (2): 31–47.

Currie, Janet, and W. Bentley MacLeod. 2017. "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians." *Journal of Labor Economics* 35 (1): 18977.

Dee, Thomas S., and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30 (3): 418–46.

Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1 (3): 111–34.

Deming, David J. 2017. "The Growing Importance of Social Skills in the Labor Market." *Quarterly Journal of Economics* 132 (4): 1593–640.

Deming, David J. 2022. "Replication data for: Four Facts about Human Capital." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E170341V1.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 2016. "School Accountability, Postsecondary Attainment, and Earnings." *Review of Economics and Statistics* 98 (5): 848–62.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2018. "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia." *Quarterly Journal of Economics* 133 (2): 993–1039.

Devereux, Kevin. 2021. "Returns to Teamwork and Professional Networks: Evidence from Economic Research." UCD Centre for Economic Research Working Paper WP21/01.

Digest of Education Statistics. 2019. "NCES 2021-009." National Center for Education Statistics, Institute of Education Sciences, US Department of Education. https://nces.ed.gov/programs/digest/.

Dimitriadis, Stefan, and Rembrand Koning. 2020. "Social Skills Improve Business Performance: Evidence from a Randomized Control Trial with Entrepreneurs in Togo." Unpublished.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2010. "Are Risk Aversion and Impatience Related to Cognitive Ability?" *American Economic Review* 100 (3): 1238–60.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2018. "On the Relationship between Cognitive Ability and Risk Preference." *Journal of Economic Perspectives* 32 (2): 115–34.

Duckworth, Angela L., Jamie L. Taxer, Lauren Eskreis-Winkler, Brian M. Galla, and James J. Gross. 2019. "Self-Control and Academic Achievement." *Annual Review of Psychology* 70: 373–99.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813.

Edin, Per-Anders, Peter Fredriksson, Martin Nybom, and Björn Öckert. 2022. "The Rising Return to Noncognitive Skill." *American Economic Journal: Applied Economics* 14 (2): 78–100.

Fe, Eduardo, David Gill, and Victoria L. Prowse. Forthcoming. "Cognitive Skills, Strategic Sophistication, and Life Outcomes." *Journal of Political Economy***.**

Figlio, David, Cassandra Hart, and Krzysztof Karbownik. 2020. "Effects of Scaling Up Private School Choice Programs on Public School Students." NBER Working Paper w26758.

Filmer, Deon, James Habyarimana, and Shwetlena Sabarwal. 2020. "Teacher Performance-Based Incentives and Learning Inequality." World Bank Policy Research Working Paper 9382.

Finkelstein, Amy, and Nathaniel Hendren. 2020. "Welfare Analysis Meets Causal Inference." *Journal of Economic Perspectives* 34 (4): 146–67.

Fryer, R. G. 2017. "The Production of Human Capital in Developed Countries: Evidence From 196 Randomized Field Experiments." In *Handbook of Economic Field Experiments*, Vol. 2, edited by Abhijit Vinayak Banerjee and Esther Duflo, 95–322. Amsterdam: North-Holland.

Fryer, Roland G. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Journal of Labor Economics* 31 (2): 373–407.

Fryer, Roland G., and Meghan Howard-Noveck. 2020. "High-Dosage Tutoring and Reading Achievement: Evidence from New York City." *Journal of Labor Economics* 38 (2): 421–52.

Fudenberg, Drew, and David K. Levine. 2006. "A Dual-Self Model of Impulse Control." *American Economic Review* 96 (5): 1449–76.

Garicano, Luis, and Esteban Rossi-Hansberg. 2006. "Organization and Inequality in a Knowledge Economy." *Quarterly Journal of Economics* 121 (4): 1383–1435.

Gennaioli, Nicola, and Andrei Shleifer. 2010. "What Comes to Mind." *Quarterly Journal of Economics* 125 (4): 1399–1433.

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2013. "Human Capital and Regional Development." *Quarterly Journal of Economics* 128 (1): 105–164.

Gill, David, and Victoria Prowse. 2016. "Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-*k* Analysis." *Journal of Political Economy* 124 (6): 1619–76.

Goldin, Claudia, and Lawrence F. Katz. 2020. "The Incubator of Human Capital: The NBER and the Rise of the Human Capital Paradigm." NBER Working Paper 26909.

Gray-Lobe, Guthrie, Parag A. Pathak, and Christopher R. Walters. 2021. "The Long-Term Effects of Universal Preschool in Boston." NBER Working Paper 28756.

Griliches, Zvi. 1977. "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica* 45 (1): 1–22.

Gunderson, Morley, and Philip Oreopoulos. 2020. "Returns to Education in Developed Countries." In *The Economics of Education*, edited by Steve Bradley and Colin Green, 39–51. 2nd ed. London: Academic Press.

Hall, Robert E., and Charles I. Jones. 1999. "Why Do Some Countries Produce So Much More Output Per Worker than Others?" *Quarterly Journal of Economics* 114 (1): 83–116.

Hansen, Stephen, Tejas Ramdas, Raffaella Sadun, and Joe Fuller. 2021. "The Demand for Executive Skills." NBER Working Paper 28959.

Hanushek, Eric A. 2003. "The Failure of Input-Based Schooling Policies." *Economic Journal* 113 (485): F64–98.

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312 (5782): 1900–1902.

Heckman, James J., and Pedro Carneiro. 2005. "Human Capital Policy." In *Inequality in America: What Role for Human Capital Policies?*, 2nd edition, edited by James J. Heckman and Alan B. Krueger. Cambridge, MA: The MIT Press.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103 (6): 2052–86.

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack. 2017. "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago." *Quarterly Journal of Economics* 132 (1): 1–54.

Hendren, Nathaniel, and Ben Sprung-Keyser. 2020. "A Unified Welfare Analysis of Government Policies." *Quarterly Journal of Economics* 135 (3): 1209–318.

Hendricks, Lutz, Christopher Herrington, and Todd Schoellman. 2021. "College Quality and Attendance Patterns: A Long-Run View." *American Economic Journal: Macroeconomics* 13 (1): 184–215.

Hendricks, Lutz, and Todd Schoellman. 2018. "Human Capital and Development Accounting: New Evidence from Wage Gains at Migration." *Quarterly Journal of Economics* 133 (2): 665–700.

Hendricks, Lutz, and Todd Schoellman. Forthcoming. "Skilled Labor Productivity and Cross-Country Income Differences." *American Economic Journal: Macroeconomics.*

Hoekstra, Mark. 2009. "The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach." *Review of Economics and Statistics* 91 (4): 717–24.

Hoffmann, Florian, David S. Lee, and Thomas Lemieux. 2020. "Growing Income Inequality in the United States and Other Advanced Economies." *Journal of Economic Perspectives* 34 (4): 52–78.

Hoffman, Mitchell, and Steven Tadelis. 2021. "People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis." *Journal of Political Economy* 129 (1): 243–85.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, & Organization* 7: 24–52.

Jackson, C. Kirabo. 2020. "Does School Spending Matter? The New Literature on an Old Question." In *Confronting Inequality: How Policies and Practices Shape Children's Opportunities*, edited by L. Tach, R. Dunifon, and D. L. Miller, 165–86. APA Bronfenbrenner Series on the Ecology of Human

Development. Washington, DC: American Psychological Association.

Jackson, C. Kirabo, Rucker C. Johnson, and Claudia Persico. 2016. "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms." *Quarterly Journal of Economics* 131 (1): 157–218.

Jackson, C. Kirabo, and Claire Mackevicius. 2021. "The Distribution of School Spending Impacts." NBER Working Paper 28517.

Jackson, C. Kirabo, Shanette C. Porter, John Q. Easton, Alyssa Blanchard, and Sebastián Kiguel. 2020. "School Effects on Socioemotional Development, School-Based Arrests, and Educational Attainment." *American Economic Review: Insights* 2 (4): 491–508.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118 (3): 843–77.

Jones, Charles I., and Paul M. Romer. 2010. "The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital." *American Economic Journal: Macroeconomics* 2 (1): 224–45.

Kaldor, Nicholas. 1961. "Capital Accumulation and Economic Growth." In *The Theory of Capital: Proceedings of a Conference Held by the International Economic Association*, edited by F. A. Lutz and D. C. Hague, 177–222. International Economic Association Series. London: Palgrave Macmillan.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan. 2015. "Self-Control at Work." *Journal of Political Economy* 123 (6): 1227–77.

Kerwin, Jason T., and Rebecca L. Thornton. 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *Review of Economics and Statistics* 103 (2): 251–64.

Khanna, Gaurav. 2021. "Large-Scale Education Reform in General Equilibrium: Regression Discontinuity Evidence from India." Unpublished.

Lange, Fabian. 2007. "The Speed of Employer Learning." *Journal of Labor Economics* 25 (1): 1–35.

Lee, Jong-Wha, and Hanol Lee. 2016. "Human Capital in the Long Run." *Journal of Development Economics* 122 (C): 147–69.

Lemieux, Thomas. 2006. "The 'Mincer Equation' Thirty Years After Schooling, Experience, and Earnings." In *Jacob Mincer: A Pioneer of Modern Labor Economics*, edited by Shoshana Grossbard, 127–45. New York: Springer.

Levine, Ross, and Yona Rubinstein. 2017. "Smart and Illicit: Who Becomes an Entrepreneur and Do They Earn More?" *Quarterly Journal of Economics* 132 (2): 963–1018.

Lindqvist, Erik, and Roine Vestman. 2011. "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment." *American Economic Journal: Applied Economics* 3 (1): 101–28.

Loyalka, Prashant, Anna Popova, Guirong Li, and Zhaolei Shi. 2019. "Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program." *American Economic Journal: Applied Economics* 11 (3): 128–54.

Ludwig, Jens, and Douglas L. Miller. 2007. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." *Quarterly Journal of Economics* 122 (1): 159–208.

Lührmann, Melanie, Marta Serra-Garcia, and Joachim Winter. 2018. "The Impact of Financial Education on Adolescents' Intertemporal Choices." *American Economic Journal: Economic Policy* 10 (3): 309–32.

Maćkowiak, Bartosz, Filip Matějka, and Mirko Wiederholt. 2021. "Rational Inattention: A Review." European Central Bank Working Paper 2021/2570.

Malamud, Ofer, and Cristian Pop-Eleches. 2011. "Home Computer Use and the Development of Human Capital." *Quarterly Journal of Economics* 126 (2): 987–1027.

Malamud, Ofer, Cristian Pop-Eleches, and Miguel Urquiola. 2016. "Interactions Between Family and School Environments: Evidence on Dynamic Complementarities?" NBER Working Paper 22112.

Mankiw, N. Gregory, David Romer, and David N. Weil. 1992. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics* 107 (2): 407–37.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *Quarterly Journal of Economics* 134 (3): 1627–73.

Mincer, Jacob A. 1974. "Schooling, Experience, and Earnings." Cambridge, MA: National Bureau of Economic Research.

Moretti, Enrico. 2004. "Estimating the Social Return to Higher Education: Evidence from Longitudinal and Repeated Cross-Sectional Data." *Journal of Econometrics* 121 (1-2): 175–212.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education?

Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review* 109 (4): 1426–60.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1): 39–77.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *Quarterly Journal of Economics* 130 (3): 1011–66.

National Institute for Early Education Research. 2021. *The States of Preschool 2020*. New Brunswick: Rugers Graduate School of Education.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92 (2): 263–83.

O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review* 89 (1): 103–24.

Oettl, Alexander. 2012. "Reconceptualizing Stars: Scientist Helpfulness and Peer Performance." *Management Science* 58 (6): 1122–40.

O*NET OnLine. 2022. National Center for O*NET Development. www.onetonline.org/ (accessed June 3, 2022).

Papageorge, Nicholas W., Victor Ronda, and Yu Zheng. 2019. "The Economic Value of Breaking Bad: Misbehavior, Schooling and the Labor Market." CEPR Discussion Paper DP14226.

Patrinos, Harry Anthony, and George Psacharopoulos. 2020. 2nd ed. "Returns to Education in Developing Countries." In *The Economics of Education*, edited by Steve Bradley and Colin Green, 53–64. London: Academic Press.

Pinkham, Amy E., David L. Penn, Michael F. Green, Benjamin Buck, Kristin Healey, and Philip D Harvey. 2014. "The Social Cognition Psychometric Evaluation Study: Results of the Expert Survey and RAND Panel." *Schizophrenia Bulletin* 40 (4): 813–23.

Rea, David, and Tony Burton. 2020. "New Evidence on the Heckman Curve." *Journal of Economic Surveys* 34 (2): 241–62.

Roser, Max, and Esteban Ortiz-Ospina. 2016. "Financing Education." Our World in Data. https://ourworldindata.org/financing-education (accessed June 3, 2022).

Roser, Max, and Esteban Ortiz-Ospina. 2017. "Pre-Primary Education." Our World in Data. https://ourworldindata.org/pre-primary-education (accessed June 3, 2022).

Rossi, Federico. 2018. "Human Capital and Macro-Economic Development: A Review of the Evidence." World Bank Policy Research Working Paper 8650.

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–90.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2019. "Capitalists in the Twenty-First Century." *Quarterly Journal of Economics* 134 (4): 1675–745.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics* 70 (1): 65–94.

Song, Jae, David J. Price, Fatih Guvenen, Nicholas Bloom, and Till von Wachter. 2019. "Firming Up Inequality." *Quarterly Journal of Economics* 134 (1): 1–50.

Sorrenti, Giuseppe, Ulf Zölitz, Denis Ribeaud, and Manuel Eisner. 2020. "The Causal Impact of Socio-Emotional Skills Training on Educational Success." University of Zurich Department of Economics Working Paper 343.

Spence, Michael. 1974. *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*. Cambridge, MA: Harvard University Press.

Springer, Matthew G., John F. Pane, Vi-Nhuan Le, Daniel F. McCaffrey, Susan Freeman Burns, Laura S. Hamilton, and Brian Stecher. 2012. "Team Pay for Performance: Experimental Evidence from the Round Rock Pilot Project on Team Incentives." *Educational Evaluation and Policy Analysis* 34 (4): 367–90.

Sunde, Uwe, Thomas Dohmen, Benjamin Enke, Armin Falk, David Huffman, and Gerrit Meyerheim. 2021. "Patience and Comparative Development." Unpublished.

Weidmann, Ben, and David J. Deming. 2021. "Team Players: How Social Skills Improve Team Performance." *Econometrica* 89 (6): 2637–57.

West, Martin R., Matthew A. Kraft, Amy S. Finn, Rebecca E. Martin, Angela L. Duckworth, Christopher F. O. Gabrieli, and John D. E. Gabrieli. 2016. "Promise and Paradox: Measuring Students' Non-Cognitive Skills and the Impact of Schooling." *Educational Evaluation and Policy Analysis* 38 (1): 148–70.

World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. Washington, DC: World Bank.

Wüst, Miriam. 2022. "Universal Early-Life Health Policies in the Nordic Countries." *Journal of Economic Perspectives* 36 (2): 175–98.

Yeager, David Scott, and Carol S. Dweck. 2012. "Mindsets That Promote Resilience: When Students Believe That Personal Characteristics Can Be Developed." *Educational Psychologist* 47 (4): 302–14.

Yeager, David S., Paul Hanselman, Gregory M. Walton, Jared S. Murray, Robert Crosnoe, Chandra Muller, Elizabeth Tipton, et al. 2019. "A National Experiment Reveals Where a Growth Mindset Improves Achievement." *Nature* 573 (7774): 364–69.

Zimmerman, Seth D. 2014. "The Returns to College Admission for Academically Marginal Students." *Journal of Labor Economics* 32 (4): 711–54.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 103–130

Measuring Human Capital

Katharine G. Abraham and Justine Mallatt

In describing the role of "fixed capital" in an economy, Adam Smith (1776, Book II, Ch. 1) considered four categories. The first three were "machines and instruments of trade," "profitable buildings," and "improvements of land." The fourth type was what economists now refer to as "human capital," which Smith described as consisting

> . . . of the acquired and useful abilities of all the inhabitants or members of the society. The acquisition of such talents, by the maintenance of the acquirer during his education, study, or apprenticeship, always costs a real expence, which is a capital fixed and realized, as it were, in his person. Those talents, as they make a part of his fortune, so do they likewise of that of the society to which he belongs. The improved dexterity of a workman may be considered in the same light as a machine or instrument of trade which facilitates and abridges labour, and which, though it costs a certain expence, repays that expence with a profit.

■ *Katharine G. Abraham is a Distinguished University Professor, University of Maryland, College Park, Maryland. She is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts, and a Research Fellow, IZA Institute of Labor Economics, Bonn, Germany. Justine Mallatt is a Research Economist, US Bureau of Economic Analysis, Washington, DC. Their email addresses are kabraham@umd.edu and Justine.Mallatt@bea.gov.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.103.

On this point, the national income and product accounts have not yet caught up with Adam Smith. The national accounts developed by Simon Kuznets and others in the 1930s and 1940s treated only investments in physical capital as additions to the capital stock. Conceptually, Kuznets recognized that this decision resulted in the omission of important investments in the nation's productive capacity. Kuznets (1961, p. 390) commented:

> [F]or many purposes—particularly the study of economic growth over long periods and among widely different societies—the concept of capital and capital formation should be broadened to include investment in the health, education, and training of the population itself, that is, investment in human beings.

While believing that, in concept, human capital should be measured, Kuznets defended its omission from the accounts on two practical grounds: first, that measuring human capital investments would be difficult; and second, that it would be hard to distinguish activities undertaken for the purpose of adding to productive capacity from those undertaken for enjoyment.

There are many reasons to want to measure investments in human capital and the resulting stock of human capital. The development of human capital is central to modern theories of economic growth (for example, Lucas 1988; Romer 1990; Barro and Sala-i-Martin 1995). Understanding how investments in the skills and abilities of the population contribute to differences in economic activity over time and across countries requires good measures of the resulting human capital and the services it provides. Failure to recognize net additions to the capital stock in the form of investments in human capital could lead to erroneous conclusions about the evolution of a country's productive capacity. The fact that spending on education and health care represents a large share of many governments' budgets also means there is considerable interest among government officials in tracking the resources devoted to those activities and the value they generate.

Investment in human capital—the skills and experience possessed by an individual or population viewed in terms of their productive value—may take many forms (Abraham and Mackie 2005). The time that parents spend with their children during the early childhood years can be considered an investment in the development of the children's cognitive, emotional, and social abilities. Formal education, from the primary grades through college and postgraduate studies, represents a further investment in the development of students' capacities. After leaving school, individuals may engage in structured training or less formal learning on the job. More broadly, one can consider medical care, diet, and exercise forms of investment in human capital. Although there have been efforts to measure investments in human capital writ broadly, investments in formal education have been of particular interest. Our primary focus, too, will be on the measurement of human capital attributable to formal education, though we also will touch on the development of more comprehensive human capital measures.

Three Approaches to Measuring Human Capital

In the US national income and product accounts, the amounts that households spend on nursery school, elementary and secondary education, vocational education, and higher education appear as part of personal consumption expenditures. The costs of the education services that nonprofit schools and colleges provide to households, over and above the revenues received in the form of tuition and fees paid for those services, also appear as part of personal consumption expenditures. The accounts record government spending on education, net of tuition revenues, separately from spending by households and nonprofits as a component of government consumption.[1]

Personal consumption spending and government spending on education as recorded in the existing accounts can be added together to produce a measure of overall education spending. Because the accounts do not consider the time that students and their parents spend on their schooling, however, that measure will understate the resources devoted to formal education. Further, because the accounts treat education spending as consumption rather than investment, they provide no information on the value of the stock of human capital attributable to investments in education or on changes in the value of that stock over time.

Three broad approaches have been taken to measuring investments in formal education and the resulting human capital stock: the indicator approach, the cost approach, and the income approach (Le, Gibson, and Oxley 2005; Jones and Fender 2011; UNECE 2016). The indicator approach attempts to capture a country's investments in human capital using measures such as school enrollment, average years of schooling, or adult literacy. The cost approach values investments in education-related human capital based on education spending. The income approach values these investments by looking forward to the increment to expected future earnings attributable to current school enrollments and calculating the present value of those added earnings. Table 1 briefly summarizes the three approaches, gives examples of studies using each of them, indicates the data required for implementing the approaches, and outlines some pros and cons for each approach. In the following pages, we discuss the indicator, the cost, and the income approach in turn.

The Indicator Approach to Measuring Human Capital

Of the three approaches to measuring investments in human capital generally and education capital specifically, the indicator approach is the most straightforward. Indicators commonly are related either to the flow of investments in education capital (as measured by, say, school enrollments) or to the stock of education capital

[1] The costs incurred by nonprofit educational institutions and government incorporate estimates of implicit spending in the form of depreciation of their physical plant and equipment. Bureau of Economic Analysis (2021) provides details on the treatment of education spending in the national accounts.

Table 1
Approaches to the Measurement of Human Capital

Approach	*Examples of Relevant Studies*	*Data Requirements*	*Pros and Cons*
Indicator: Measure or measures indicative of a country's investment in or stock of human capital; if multiple measures, weighted to form an index.	Barro and Lee 1993, 2001, 2013, 2021 (average years of schooling) Kraay 2019, World Bank 2020 (World Bank Human Capital Index; considers expected years of schooling, test scores, prevalence of stunted growth, and child and adult survival rates) Samans et al. 2017 (World Economic Forum Global Human Capital Index; considers measures that include school enrollment, educational attainment, literacy, labor force participation, and skill mix of employment)	Survey, census, or administrative data for chosen metric(s) that are consistent across countries and over time	(+) Relatively straightforward to construct and explain (-) Schooling measure(s) may mean different things in different contexts (-) Weights for indicators that combine multiple measures can be arbitrary (-) Not compatible with national accounts or measures of other types of capital
Cost: Current gross investment equals direct spending plus estimated value of unpaid time devoted to human capital development; stock equals sum of appropriately depreciated past investments.	Kendrick 1976 (expanded accounts encompass investments in child rearing, formal education, training, health, and geographic mobility) Eisner 1978, 1985, 1989 (Total Incomes System of Accounts encompasses investments in formal education, training, and health) Gu and Wong 2015 (recent cost-based estimates of investments in formal education in Canada; do not report stock estimates)	School enrollment by age, sex, and type of schooling (e.g., grade level); participant numbers for other human capital investments Direct spending for formal education, training, and other human capital investments Value for time devoted to human capital investment (e.g., student time in formal education, employee time in training)	(+) Monetary measure suitable for integration into national accounts and compatible with measures of other types of capital (-) Relatively demanding data requirements, especially for investments other than in formal education (-) Sensitive to assumptions about value of nominal spending in different periods and rate at which investments depreciate (-) Captures resources devoted to formal schooling and (if applicable) other human capital investments, not necessarily the productive value of that spending
Income: Current gross investment equals year-over-year additions to present value of future labor income; stock equals present value of current population's future labor incomes.	Jorgenson and Fraumeni 1989, 1992a, 1992b (estimates of investment through formal education and additions to population, and of value of stock of human capital) Christian 2010, 2014, 2017; Fraumeni and Christian 2019 (update and extend earlier Jorgenson and Fraumeni work)	School enrollment by age, sex, and type of schooling (e.g., grade level) Population by age, sex, and educational attainment Earnings by age, sex, and educational attainment Mortality rates by age and sex	(+) Monetary measure suitable for integration into national accounts and compatible with measures of other types of capital (-) Relatively demanding data requirements (-) Sensitive to assumptions regarding future growth in earnings, appropriate discount rate for future earnings, and, for formal education, how not completing a year of schooling affects later educational attainment

(as proxied by, say, adult educational attainment or adult literacy). The indicator approach provides a relatively parsimonious way to compare investments in human capital across countries.

Perhaps the best-known indicator dataset has been developed by Barro and Lee (1993, 2001, 2013, 2021). The latest version of the Barro and Lee dataset contains information on educational attainment for 146 countries. It reports the share of the population with each of seven levels of education—no formal education, incomplete primary, complete primary, lower secondary, upper secondary, incomplete tertiary, and complete tertiary—by five-year age intervals over the period from 1950 through 2015. The dataset also contains measures of mean years of schooling. The Barro-Lee measures are based primarily on national census data and, in some cases, have suffered from apparent inconsistencies over time. Seeking to resolve these inconsistencies, several teams of researchers have proposed alternative educational attainment series as substitutes for the Barro-Lee schooling measures (for example, Cohen and Soto 2007; Cohen and Leker 2014; de la Fuente and Doménech 2015; Goujon et al. 2016).

Other prominent indicator-based human capital measures are weighted indexes based on multiple underlying components that encompass more than just formal education. The most recent version of the World Bank's Human Capital Index covers 174 countries. The dimensions incorporated in this index include the probability of survival to age five; expected years of school; harmonized test scores; the fraction of children under age five whose growth is not stunted; and adult survival rates (Kraay 2019; World Bank 2020). As another example, the World Economic Forum's Global Human Capital Index, last published in 2017 for 130 countries, incorporates an even larger set of indicators. These capture not only what its authors term development (formal education of the next generation workforce and upskilling of the current workforce) and capacity (level of formal education resulting from past investments), but also knowhow (breadth and depth of specialized skills in use at work) and deployment (skills application and accumulation among the adult population) (Samans et al. 2017).

One challenge in constructing indexes like the Human Capital Index and Global Human Capital Index is the selection of weights for the various index components. Rather than using the index, researchers may choose to employ the underlying index components.

The indicator approach has been useful. Data on school enrollments and educational attainment not only are valuable in themselves, but also are necessary inputs to the full development of the other two approaches to the measurement of education capital—the cost and income approaches. Various studies have used indicators based on educational attainment in empirical analyses of economic growth. Some of these studies simply include mean years of schooling in cross-country growth regressions, effectively treating the productive value of additional years of schooling as a constant. Others make use of information on years of schooling, but allow the returns to

education to vary with the educational attainment of the population or over time (Botev et al. 2019).[2]

The indicator approach also has limitations. While a measure such as mean years of schooling may help with understanding differences in productivity over time or across countries, depending on the content and quality of the education provided, a year of schooling may mean different things at different times and in different countries. This problem can be addressed to some extent by using proxies such as student-teacher ratios or test scores to adjust for varying educational quality (Fraumeni et al. 2009; UNECE 2016; Kraay 2019). In addition, the value of the human capital produced through formal education may depend on other factors, like a country's institutions and social infrastructure (Hall and Jones 1999; Caselli and Ciccone 2019).

On their own, measures of educational attainment cannot provide answers regarding the value-added of the education sector. More generally, the indicator approach is not designed for compatability with the treatment of other types of investment in the national accounts. Monetary measures constructed using methods compatible with national accounting practices, such as the cost-based and income-based measures we consider next, are more appropriate for any analysis that considers human capital investment in the context of investment and capital accumulation more broadly.

The Cost Approach to Measuring Human Capital

Tracking changes in nominal education spending over time is relatively straightforward. Translating a data series for nominal spending on education into a real spending series, and then using those data to construct estimates of capital depreciation and the stock of education capital, is considerably more challenging. Data limitations make carrying out these tasks for other types of human capital investment even more difficult. Perhaps for these reasons, relatively few researchers have adopted the cost-based approach to measuring human capital. During the 1970s and 1980s, Kendrick (1976) and Eisner (1978, 1985, 1989) developed expanded economic accounts that incorporated human capital investment based on a cost approach. Their efforts were ambitious, encompassing not only investments in formal education but also investments in job training, health, and, in the case of the Kendrick estimates, geographic mobility and child rearing. More recently, Gu and Wong (2015) have developed both cost- and income-based estimates of investments in formal education for Canada.

A recent international task force operating under the auspices of the United Nations Economic Commission for Europe has developed guidelines for satellite accounts that would systematically compile information on the costs of education and training. These satellite accounts would provide much of the information needed to construct cost-based measures of real education investment and the stock

[2]Incorporating information on the returns to education creates some similarities between these approaches and the income approach discussed below.

of education capital, should a national statistical office wish to do so (UNECE 2020). Here, we discuss several of the challenges in developing such measures: converting nominal expenditures on education to real terms, estimating a capital stock based on past expenditures, and the issues posed by the valuation of time spent by children and parents in education and by immigration and emigration.

Understanding how investments in education capital have evolved over time and producing measures of the current capital stock requires information on *real* education spending, as opposed to nominal spending. The standard approach to converting from nominal to real spending is to use an index of output prices to adjust spending amounts for the effects of inflation. Because governments do not sell education services at market prices, however, that approach will not work in this area. An alternative approach for estimating the real value of government output is to deflate spending by an index of the prices of *inputs* for education—like teacher salaries—but this has the drawback of assuming that the technology for transforming education inputs into education outputs does not change over time. For government output that has a market counterpart, deflators could be constructed using private sector prices, but even when the government and the private sector appear to provide similar services, they may not be truly comparable.

National accounting experts who have considered how best to measure nonmarket output, including government output, generally have concluded that the best approach is to use a quantity index for apportioning nominal changes in spending into the piece that presents real output change versus the piece that is due to changes in prices (European Commission et al. 2009; UNECE 2016). In the case of education, the challenge is to produce a quantity index of real output that can be tracked over time.

A simple way to construct a quantity index for education would be as follows. Start with the number of students educated in a base period. Divide these students into "types," for example, by grade. Get information on the share of education spending going to each type of student in the base year. For each future year, get data on the number of students of each type and use that information to construct a "quantity relative" equal to the number of that type in the later period divided by the number in the base period. Use the cost weights from the base period to sum up these quantity relatives for the different types of students. The result will be a base-weighted (or Laspeyres) quantity index.[3] The changes in nominal spending over time then can be divided into the real change in output (the portion of the change accounted for by the change in the quantity index) and the change in price (the piece that is left over).

These calculations implicitly assume that the quality of education for students within a particular group does not change over time, a necessary assumption so

[3] Index number formulas such as the chained Fisher or chained Tornqvist formula generally are preferred to the Laspeyres formula, but quantity indexes constructed using these formulas could be used in the same fashion to estimate real expenditures by period. For a discussion of alternative index formulas and their properties, see Diewert (2021).

that cost-share-weighted sums of the quantity relatives for the different groups of students can be used to measure education output in the later period compared to the earlier period. As mentioned in discussing the indicator approach, it may be possible to improve the measures of real output over time by incorporating proxies for the quality of education into the analysis, although with a tradeoff of additional complexity.

Given data on past real investments in capital, one can use the "perpetual inventory method" to develop an estimate of the current value of the resulting stock of capital. Except for automobiles, which are valued directly, this is the approach used in the existing national income and product accounts for valuing the current physical capital stock (Katz 2015). The basic idea is that the change in the value of the capital stock from one year to the next equals new investment spending minus an adjustment for any year-over-year decline in the value of the previously existing capital stock. The key question is how much spending on capital in earlier periods contributes to the stock of capital in the present.

Physical capital depreciates with age both because it becomes less efficient (for example, because it requires more maintenance downtime) and because its remaining useful lifetime is shorter. Sales of used assets provide direct evidence on the depreciation of physical capital over time. Similarly, human capital may depreciate both because of changes in the value of the skills a person possesses (for example, skills acquired in school may become rusty over time) and because expected remaining lifetimes become shorter as people age. In contrast to physical capital, though, no direct evidence is available for quantifying how human capital depreciates. Past estimates of the stock of human capital based on the cost approach have made differing assumptions about depreciation profiles, but there is little empirical basis for choosing among them. Eisner (1978), for example, assumed straight-line depreciation, that is, that a human capital asset with a useful life of T years loses 1/T of its initial value each year. Kendrick (1976) assumed "double-declining balance depreciation," meaning geometric depreciation at a rate equal to twice that implied by straight-line depreciation in the first year of the asset's life, switching over to straight-line depreciation at the point when that became larger than the depreciation implied by the double-declining balance method. This difference in the choice of depreciation method explains why Eisner estimates larger values for net investment (gross investment minus depreciation) in human capital than Kendrick.

Two additional points about the cost-based approach to valuing the human capital created through formal education are worth noting here. First, as already remarked, although considerable information on education spending is available from the existing national income and product accounts, they omit the value of the hours that students spend in school or studying, together with the value of the hours spent by parents in supporting the students. The value of this unpaid time is an important part of the true cost of formal education.

The appropriate valuation for the time students devote to their own education is their opportunity wage—the amount that they could have earned had they been working rather than in school. Because the services provided by parents could be

performed by someone else, the right wage for valuing that time is a replacement wage—what it would have cost to hire someone else to do the same work—rather than an opportunity wage (Abraham and Mackie 2005).

In the United States, reasonable estimates of the hours students devote to schooling can be constructed using data on school enrollments by grade level, attendance rates, and academic calendars compiled by the National Center for Education Statistics (NCES). Compulsory schooling and child labor laws typically prevent younger children from working for pay, so it is reasonable to set the opportunity cost of younger children's time to zero. For older students, however, the earnings foregone by remaining in school are a significant part of the cost of their education. Beginning in 2003, estimates of the time that parents spend on activities related to children's education are available from the American Time Use Survey. As discussed later, the hours that parents devote to children's formal education—and thus the value of that parent time—are modest relative to the value of the time students devote to their own education.

An additional complication in valuing the time that students devote to their education is that, in addition to contributing to a person's human capital, education also may be something that people enjoy and thus a form of consumption. Conceptually, to the extent that being in school is more enjoyable than working at a job, some portion of the opportunity cost associated with the time students devote to formal education should be treated as consumption rather than investment. On the other hand, some students may find being in school particularly unpleasant. In that case, the adjustment should go in the other direction, implying a true cost of education that is higher than estimated based on direct expenditures and the opportunity cost of students' time. One interesting recent study suggests that students derive significant positive consumption value from being in school (Gong et al. 2021), but this is very much an open area for further research.

A final comment about the cost approach for measuring human capital is that estimates of the stock of education capital based on past education spending do not account for the effects of immigration and emigration. Most immigrants arrive as adults, meaning that, on arrival, they embody a significant amount of human capital acquired elsewhere. In 2019, 13.7 percent of the resident population of the United States had been born somewhere else (Levesque and Batalova 2022). A full assessment of how a country's stock of human capital evolves would need to account for the additions to the stock through immigration and losses through emigration.

The Income Approach to Measuring Human Capital

In a frictionless market that operates with complete information, the price that someone should be willing to pay for a marketable asset is equal to the present value of the future returns that asset will generate. In a series of seminal papers, Jorgenson and Fraumeni (1989, 1992a, 1992b) adapt the spirit of this approach to value investments in formal education (and other forms of human capital)—that is, to calculate the stock of human capital by estimating the present value of the future returns that workers will generate.

Using the Jorgenson and Fraumeni approach to estimate the value of the stock of human capital requires data on the number of people in the population by age, sex, and level of education. The calculations also require estimated earnings for each age/sex/education cell, together with the probabilities of survival from one year to the next. Jorgenson and Fraumeni begin with current figures on the earnings of people in different age/sex/education cells and assume that the overall level of earnings will grow by some percentage g each year, but that the relative earnings of people in the different age/sex/education cells will not change. Future earnings are discounted to capture present values. Here, we first describe the basics of the Jorgenson and Fraumeni income approach, and then discuss three challenges in its implementation.[4]

As a starting point to determining the expected present value of future earnings for people of a given age, sex, and level of education, Jorgenson and Fraumeni begin by calculating the present value of lifetime earnings for the oldest individuals in their data set and work recursively backwards. Suppose that the oldest working people are age 75. The present value of market income for someone in this group is just equal to market income at age 75. Now consider the present value of lifetime earnings for a person age 74. This equals current earnings as of age 74 plus the expected present value of future earnings as of age 75. Jorgenson and Fraumeni continue working backwards in the same fashion to younger age groups. In this way, they estimate expected future earnings for everyone in the population, differentiated by age, sex, and level of education.

Investments in formal education are valued based on projections of the amount they will add to future earnings. The total value of the human capital stock may grow from one year to the next due not only to formal education, but also due to births or in-migration. Conversely, the value of total human capital can decline from one year to the next due to aging (which reduces years of anticipated future earnings for the existing population), death, and outmigration.[5]

In contrast to the cost approach, the income approach does not require explicit assumptions about the rate of depreciation of human capital, as that can be backed out of the calculations by looking at how the expected present value of earnings changes as people age. It does require other assumptions, including assumptions about the growth rate of the overall level of future earnings and the intertemporal discount rate. Typical assumptions regarding the annual growth of labor income and the discount rate for future earnings are in the range of 1 to 2 percent per year

[4] Klenow and Rodríguez-Clare (1997) and Arrow et al. (2012) lay out a different approach for calculating the present value of the stream of lifetime income attributable to investments in human capital and valuing the stock of human capital. The United Nations Environment Programme has employed this method for its Inclusive Wealth Report (Managi and Kumar 2018), but studies applying the income approach to valuing human capital more commonly have adopted the methods developed by Jorgenson and Fraumeni.

[5] The human capital stock also may be revalued from one period to the next if there are changes in projected earnings for people of given age, sex, and education.

for the growth rate of future earnings and 4 to 5 percent per year for the intertemporal discount rate.

The choice of discount rate is of particular interest in part because the discount rate that a social planner would apply may be substantially lower than that applied by individuals making decisions about investments in education. There are two reasons for this. First, because individuals cannot diversify their investment in formal education, they will correctly view such investments as risky. From the perspective of the society as a whole, however, investment in formal education is diversified across individuals and thus considerably less risky. Second, individuals often appear time-inconsistent with regard to their educational decisions, choosing lower levels of investment in education than they later wish they had made (DeGenova 1992). To overcome this time inconsistency, a social planner should place more weight on future utility than would individual decision-makers, leading to a social discount factor lower than the individual discount factor (Caplin and Leahy 2004).

Estimates of the value of investment in education are quite sensitive to choices regarding the earnings growth rate and discount rate. Jorgenson and Fraumeni (1992b) report that the value of investments in formal education, including both market and nonmarket returns, was $5.0 trillion in 1986 (in 1986 dollars) assuming an annual earnings growth rate of 2 percent and a discount rate of 4 percent. This total falls nearly by half to $2.7 trillion assuming an earnings growth rate of 1 percent and a discount rate of 6 percent. Christian (2014) reports that, in 2009, the market value of gross investment in human capital calculated using the standard Jorgenson and Fraumeni approach was $21.0 trillion (in 2009 dollars) assuming an annual earnings growth rate of 2 percent and an annual discount rate of 4 percent. Assuming instead an annual growth rate in earnings of 1 percent and an annual discount rate of 12 percent, this falls to $3.1 trillion.

A first challenge in implementing the income approach is how to value the human capital of those who have not yet completed their education. In the Jorgenson and Fraumeni calculations, persons age 35 through 75 do not enroll in school, but individuals between ages 5 and 34 may choose to acquire additional education. In describing how they project future labor income for a person with either the highest or the next-highest number of years of education, Jorgenson and Fraumeni (1992b, p. 309) explain:

> For an individual of a given age and sex enrolled in the highest level of formal schooling, which is the 17th year of school or higher, lifetime labor income is the discounted value of labor incomes for a person with 17 years or more of education. For an individual enrolled in the 16th year of school, lifetime labor income includes the discounted value of labor incomes for a person with 17 years of formal education or more, multiplied by the probability of enrolling in the 17th year of school, given enrollment in the 16th year . . . It also includes the discounted value of labor incomes for a person with 16 years of education, multiplied by one minus this probability, which is the likelihood of terminating formal schooling at 16 years.

For an individual of any given age and current schooling level, the value of investing in an additional year of schooling is treated as equal to the difference between the expected present value of labor income for a person who completes the extra year of schooling versus that for someone who does not. This includes any differences in future earnings related to the fact that those who complete the extra year of schooling are more likely than those who do not to continue on to acquire further schooling.

A difficulty with these calculations is that future school enrollments among the set of people not completing the extra year of schooling may provide a poor guide to what would have happened to the person who finished the extra year of schooling had they not done so. Consider a 17 year-old with 11 years of schooling who completes a 12th year of schooling and graduates from high school during the following year. To determine the value of that extra year of education, Jorgenson and Fraumeni would compare the projected future earnings of the 18 year-old with 12 years of schooling to the projected future earnings of an 18 year-old with 11 years of schooling. The problem is that an 18 year-old with just 11 years of schooling is someone who has fallen off track educationally. The probability of that individual continuing with their education is low. Because the people who continue on in school almost certainly differ in their ability, motivation, and other characteristics from those who drop out, however, the experiences of the dropout group may not provide a good indication of what would have happened to the person completing their 12th year of school had they failed to graduate at age 18 (Christian 2010). An alternative counterfactual for what would have happened had the 18 year-old not completed year 12 might be that the probability of their doing so is the same as for a 17 year-old with 11 years of schooling—a person who is still on track educationally.

Christian (2010) shows that assumptions about future enrollments can have a very large effect on the estimated returns to formal education. In one illustration, under the standard Jorgenson and Fraumeni counterfactual, the market component of gross investment in education had a value of $16 trillion in 2005. Under the alternative assumption that, had a person who acquired a year of education not done so, the probability of their doing so subsequently would have been the same as for a person with the same initial education who is a year younger, the market component of gross investment in education in 2005 is just $3.1 trillion.[6]

A second challenge for the income-based approach is how best to consider the benefits of human capital for individuals not engaged in market work, but who engage in enjoyable or productive non-market activities. Jorgenson and Fraumeni decide to value both market and nonmarket time. They reason that individuals will

[6]These calculations assume an annual growth rate in earnings of 2 percent and a discount rate of 4 percent. For his own estimates, Christian (2010) chooses to focus on the net return to education—comparing the projected earnings of a person of age $a+1$ with $e+1$ years of schooling to those of a person of age a with e years of schooling—rather than the gross returns. As can be seen in Figure 5 of the working paper version of his paper (Christian 2009), however, at the discount rates he assumes, this yields results very similar to calculating gross returns under the second of the counterfactual assumptions just discussed.

choose to work up to the point where the marginal return to working just equals the marginal value of time at home. They assume 10 hours per day devoted to personal maintenance activities and (at younger ages) 1,300 hours per year devoted to education by people who are in school. Then, they treat the value of non-market time as equal to the (actual or imputed) marginal after-tax wage rate.

The decision to count both the market and the nonmarket returns to education has a very large effect on income-based estimates. In the original Jorgenson and Fraumeni (1992b) analyses, the value of investment in education including both market and nonmarket returns is 2.3 to 3.2 times as large as the market component on its own, depending on the year. Similarly, in evaluating the returns to investment in education net of aging, Christian (2014) estimates total values that are roughly double the values based on market returns alone.

Even counting only the market returns to education, estimates of the value of investments in education are very large compared to investments in other assets. Jorgenson and Fraumeni (1992b), for example, report that formal education in 1986 raised the present value of the market returns to education by about $1.6 trillion dollars (in 1986 dollars). This is close to double total gross private domestic investment for the same year, based on data from the US Bureau of Economic Analysis. Including nonmarket returns, the Jorgenson and Fraumeni estimate of the value of investment in educational capital in 1986 was roughly 4.5 times as large as the official estimates for gross private domestic investment. Studies for other countries have found similar or larger multiples. Liu (2014) reports that, in a set of 10 OECD countries as of 2006, ratios of the value of the stock of human capital estimated based just on the market returns to the value of the stock of physical capital ranged from 3.7 in the Netherlands to 7.0 in the United Kingdom. Due both to discomfort with the even larger values obtained when nonmarket returns are included and, more importantly, the additional data and assumptions required to value nonmarket returns, applications of the Jorgenson and Fraumeni income-based approach often have focused only on the market returns.

A third challenge for the income-based approach is that, among those of a given age and sex, all differences in future earnings between people with different levels of education are attributed to the differences in their educational attainment. Some of these differences may in fact be a result of returns to other types of human capital investment. For example, the higher earnings of more educated people may reflect not only returns to education but also returns to larger early childhood investments (Björklund and Salvanes 2011).

Returns from on-the-job training also might bias estimates of the value of investments in formal education. After completing their schooling, highly educated workers are more likely to participate in on-the-job training than are less educated workers (for example, Bureau of Labor Statistics 1996; Eurostat 2022). Moreover, educated workers experience steeper growth in earnings with experience. It is not obvious, however, whether this leads to bias in estimates of the value of education. Conceptually, a worker entering the labor market will choose among career paths with different amounts of on-the-job training and different wage profiles. In

market equilibrium, all of the career paths available to a worker should offer the same expected present value of earnings (Becker 1964). If workers apply the same discount rate in evaluating present and future income that the analyst uses when constructing income-based estimates of investment in educational capital, then the analyst's calculations should not be affected by whether educated people also invest more in on-the-job training. If, however, workers apply a higher discount rate in deciding whether to make on-the-job training investments, then when more-educated workers nonetheless choose more training, the estimated value of acquiring additional education will be upward biased. O'Mahony and Stevens (2009) is one paper that has recognized the potential confounding of returns to education and returns to experience.

Empirical Measures of Human Capital Investments and Stocks

In this empirical discussion, we begin with a short review of cross-country evidence on human capital, then turn to a comparison of estimates based on the cost and income approaches. As already discussed, the cost approach to measuring the value of investments in education is based on the costs of producing formal education; the income approach attempts to value the resulting output. Similar to the way in which the income-based and expenditure-based estimates of gross domestic product embedded in the double-entry bookkeeping of the national income and product accounts provide a check on one another, it would be reassuring if the estimates of human capital investment based on the cost and income approaches were of similar magnitude. In practice, estimates of the value of investments in human capital based on the income approach have been far larger than estimates based on the cost approach. We discuss why the two approaches might produce such different answers and whether there is a way to reconcile them.

Cross-Country Evidence

Investigating how differences in human capital contribute to cross-country differences in economic growth requires a measure produced in a comparable fashion across countries and over time. Candidates in the literature include various indicator measures of human capital—for example, measures of years of schooling like those in the Barro-Lee dataset (Barro and Lee 2021), the World Bank Human Capital Index (HCI) (World Bank 2020), and the World Economic Forum Global Human Capital Index (GHCI) (Samans et al. 2017). Perhaps surprisingly, though there would be no conceptual barrier to producing a measure of human capital investment suitable for cross-country comparisons based on the cost approach, no such measure appears to exist. Both the World Bank and the United Nations have produced income-based measures of human capital investment.

The income-based measure of human capital developed by the World Bank (defined as the present value of current and future market incomes for the population age 15–64) uses the approach developed by Jorgenson and Fraumeni to

assign present values to individuals in different age/sex/education cells. The World Bank's 2018 Changing Wealth of Nations report contains estimates for 2014 for 141 countries developed using information from its extensive database of household surveys; market exchange rates were used to convert the country-specific numbers to US dollars (Lange, Wodon, and Carey 2018).[7] Using a somewhat different approach based on Klenow and Rodríguez-Clare (1997) and Arrow et al. (2012), the United Nations Environment Programme (UNEP) has produced alternative income-based measures of the stock of human capital. Its 2018 Inclusive Wealth Report makes use of estimates for 2014 for 140 countries.[8] These are estimates of the value of the human capital possessed by adults who are past the age normally required to complete their reported level of education based on an assumed rate of return to schooling. Although conceptually similar, the UNEP Inclusive Wealth Report estimates differ from the World Bank Changing Wealth of Nations estimates in several ways. First, in these estimates, each year of education raises the human capital that a person possesses by a fixed percentage amount. Second, the calculations make no distinctions based on the likelihood that a person will work for pay, so that the estimates capture both market and nonmarket returns to education. Third, the country-specific human capital values were converted to US dollars using purchasing power parities rather than market exchange rates (Managi and Kumar 2018).

Although these different measures have distinct underpinnings, one can ask whether they vary similarly across countries and over time. In Table 2, we report cross-country correlations for the five measures mentioned in the preceding paragraphs.[9] In addition to the two income-based measures for which we have 2014 data, the calculations use Barro-Lee data for 2015, HCI data for 2017, and GHCI data for 2017; the years were chosen to be as close together as possible given the available information.[10] To scale the income-based measures, we use the natural logarithm rather than the level of the per capita value of countries' human capital, which is similar to using percentage differences rather than absolute differences across countries in the calculations. The Table 2 correlations are Pearson correlations that represent the covariances between pairs of measures across countries, standardized by dividing by the product of the standard deviations of the two series, so the resulting number always lies between –1 and 1.

All five of the measures we examine are positively correlated with each of the others. The most closely related are the three indicator measures (the Barro-Lee

[7] These estimates built on an earlier initiative undertaken at the Organisation for Economic Cooperation and Development (Liu 2011).

[8] Barbara Fraumeni kindly shared these data with us.

[9] Liu and Fraumeni (2020) report correlations similar to those reported here for a somewhat different set of measures.

[10] HCI data for slightly fewer countries are available for 2017 (157 countries) than for 2020 (174 countries). We drop the estimated per-capita value of human capital in Slovakia in the IWR data because it is prima facie implausible (nearly 20 times as large as the estimated US value), leaving us with IWR data for 139 countries.

Table 2
Correlations across Countries for Selected Human Capital Measures

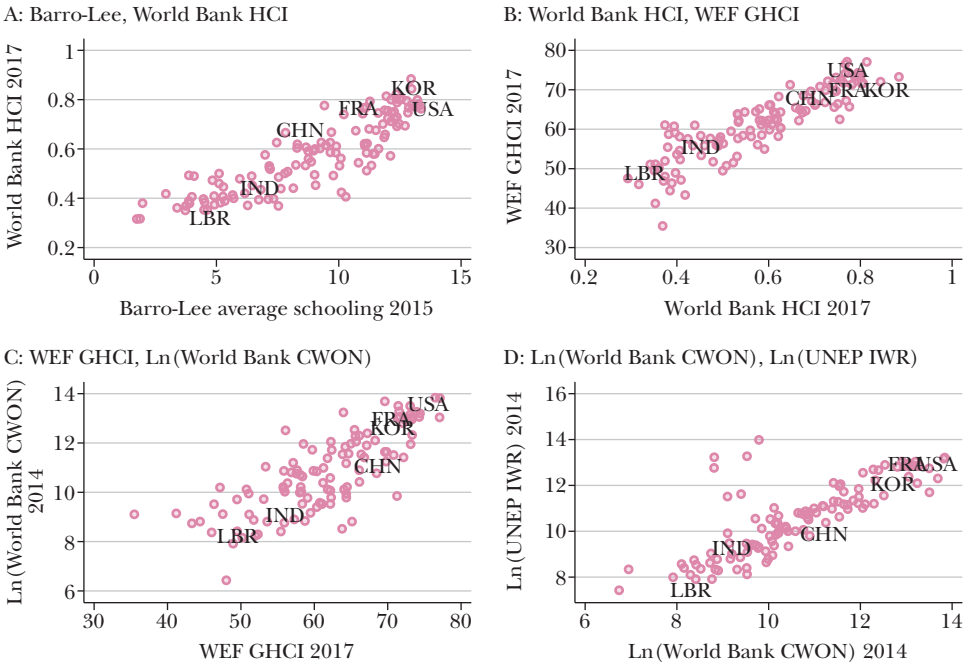
Measure	*Indicator: Barro-Lee years of schooling*	*Indicator: World Bank HCI*	*Indicator: World Economic Forum GHCI*	*Income-based: Ln(World Bank CWON)*	*Income-based: Ln(UNEP IWR)*
Indicator: Barro-Lee years of schooling	1.000 (146)	—	—	—	—
Indicator: World Bank HCI	0.872 (132)	1.000 (157)	—	—	—
Indicator: World Economic Forum GHCI	0.852 (124)	0.892 (126)	1.000 (130)	—	—
Income-based: Ln(World Bank CWON)	0.796 (122)	0.850 (132)	0.788 (117)	1.000 (141)	—
Income-based: Ln(UNEP IWR)	0.691 (138)	0.774 (130)	0.656 (123)	0.814 (122)	1.000 (139)

Source: Authors' calculations.
Note: HCI = Human Capital Index. GHCI = Global Human Capital Index. CWON = Changing Wealth of Nations. UNEP = United Nations Environment Programme. IWR = Inclusive Wealth Report. Barro-Lee data for 2015; World Bank HCI data and World Economic Forum GHCI data for 2017; World Bank CWON and UNEP IWR data for 2014. Income-based measures of human capital are ln(per capita value). Numbers in parentheses are counts of countries for which each pair of measures available. Implausible UNEP IWR value for Slovakia dropped.

measure of years of schooling, the World Bank HCI, and the World Economic Forum GHCI); each of the pairwise correlations involving these measures lies above 0.85. The UNEP Inclusive Wealth Report ln(per capita value of human capital) measure is less highly correlated with the indicator measures than the World Bank Changing Wealth of Nations ln(per capita value of human capital) measure. Perhaps surprisingly given their conceptual similarity, the correlation between the two ln(per capita value of human capital) measures is not especially high but rather lies in the middle of the pack.

Figure 1 contains scatterplots of selected pairs of measures. Panel A plots the Barro-Lee measure against the HCI; panel B, the HCI against the GHCI; panel C, the GHCI against the World Bank Changing Wealth of Nations ln(per capita value of human capital) measure; and panel D, the World Bank Changing Wealth of Nations ln(per capita value of human capital) measure against the conceptually similar UNEP Inclusive Wealth Report measure. In each pairing displayed, countries with high values on one measure tend to have high values on the other. It also is clear, though, that the relationships between the two pairs of indicator variables displayed in panels A and B are stronger than the relationships involving one or both of the ln(per capita value of human capital) measures displayed in panels C and D. Even after removing the UNEP Inclusive Wealth Report value for Slovakia as implausible on its face, there are six dots in panel D for which the UNEP Inclusive Wealth Report values lie well above the level expected based on the World

Figure 1
Relationships between Selected Pairs of Human Capital Measures



Source: Authors' calculations.
Note: HCI = Human Capital Index. GHCI = Global Human Capital Index. CWON = Changing Wealth of Nations. UNEP = United Nations Environment Programme. IWR = Inclusive Wealth Report. Barro-Lee data for 2015; World Bank HCI data and World Economic Forum GHCI data for 2017; World Bank CWON and UNEP IWR data for 2014. Income-based measures of human capital are ln(per capita value). Implausible UNEP IWR value for Slovakia dropped.

Bank Changing Wealth of Nations measure. These dots represent Cote D'Ivoire, Kyrgyzstan, Moldova, Tanzania, Turkey, and Vietnam. The UNEP Inclusive Wealth Report numbers place the per capita value of these countries' human capital well above that in other developing economies, at or above the levels for the US and other developed nations. These anomalous results suggest that, at least for 2014, the World Bank Changing Wealth of Nations measures should be preferred to the UNEP Inclusive Wealth Report measures.

Estimates of Human Capital: Investment and Stock

Applying the tools of growth accounting to human capital requires monetary measures of human capital constructed using methods more consistent with those used in the existing national income and product accounts—the cost-based and income-based measures discussed above. Discussing his cost-based capital stock estimates, Kendrick (1976, p. 19) states: "Our net capital estimates in current prices . . . approximate market values, assuming reasonably good foresight by the businessmen

[sic] who made the investment decisions." In other words, he argues, the amount that an informed individual making an asset purchase would spend should be just the anticipated present value of the returns to that asset. To the extent that similar reasoning applies to human capital, the cost and income approaches to estimating investments in human capital should give similar answers. In practice, where both are available, estimates of investment in education—and other types of human capital—using the income method based on the valuation of future returns have been much larger than corresponding estimates based on the costs of the resources devoted to these investments.

Consider the relative magnitudes of the cost-based estimates of the value of investment in education and training reported by Kendrick (1976) and the income-based estimates of investment in formal education reported by Jorgenson and Fraumeni (1992b). The estimates of human capital from these two sources, which overlap for the years from 1947 to 1969, are dated but remain the most authoritative available for comparison purposes. Kendrick's cost-based estimates are in some ways more inclusive than the Jorgenson-Fraumeni estimates. In addition to direct spending on schools and an estimate of the opportunity cost of student time, Kendrick's estimates include spending on libraries, religious education, and employee training, as well as a portion of spending on radio, television, books, and other items that are treated as having educational value. The Jorgenson and Fraumeni estimates refer strictly to the incremental returns to additional years of formal education. Despite their more restricted scope, the Jorgenson and Fraumeni estimates are 6 to 9½ times as large as the Kendrick estimates, depending on the year. Even if one looks only at the market returns to education in constructing the income-based estimates, the Jorgenson and Fraumeni results imply values for the investment in education that are 2 to 3½ times as large as those reported by Kendrick.[11]

In calculations using Canadian data, Gu and Wong (2015) report estimates of the value of investments in formal education on both a cost and an income basis for the period from 1976 through 2005. The differences they find between income-based estimates of market returns to formal education and the corresponding cost estimates are even more striking, with the former roughly 6 to 14 times as large as the latter, depending on the year (see their Figure 5).

Estimates of the total stock of human capital using cost-based versus income-based methodologies—including both education capital and human capital acquired through other types of investments—are even more different.[12] To estimate the value of the stock of human capital, Kendrick (1976) takes into account the costs of rearing individuals to the point at which they can be productive, including the value of the time their parents spent caring for them as young children, together with the

[11] The cited estimates refer to current-dollar cost-based estimates of the value of investment in education from table B–2 of Kendrick (1976); current dollar income-based estimates from table 8.6 of Jorgenson and Fraumeni (1992b); and estimates of the share of the value of investment in education accounted for by market income from table 8.11 of Jorgenson and Fraumeni (1992b).

[12] We can't compare estimates of the stock of education capital based on the two approaches because that isn't separately identified in the Jorgenson and Fraumeni numbers.

costs of food, clothing, shelter, and so on. He combines these costs with spending on health, education, and training, then applies the perpetual inventory method to the spending series to obtain stock estimates. In the alternate approach, income based-based estimates reported by Jorgenson and Fraumeni (1992b) value the future flow of income to the current population by age, sex, and level of education. In each of the years for which the estimates can be compared (1948 through 1969), the Jorgenson and Fraumeni estimates of the value of the total stock of human capital are roughly 18 times as large as the Kendrick estimates. Even if the income-based estimates are adjusted to consider only the contribution of market earnings to the value of the stock of human capital, the Jorgenson and Fraumeni income-based estimates are still 5 or 6 times as large as Kendrick's cost-based estimates.[13]

Seeking a Reconciliation

Why do estimates of human capital from income-based approaches tend to dwarf their cost-based counterparts? Our sense is that the divergence is more likely to be the result of overstatements by the income approach than understatements from the cost approach.

It is possible that past efforts using the cost-based approach have understated the full cost of education, but it seems unlikely that any understatement could be large enough to make a significant dent in the very large observed discrepancies in the two sets of estimates.

There are, however, several plausible reasons why estimates based on the income approach might overstate the value of investments in education. The income-based approach could 1) apply an intertemporal discount rate that is too low (or equivalently, an expected growth rate in future earnings that is too high); 2) overestimate the returns to education by understating the counterfactual earnings prospects for those who acquire additional education; 3) exaggerate the returns to education by valuing nonmarket time for educated workers based on their market wage; and 4) confound the returns to education with the returns to other investments in human capital (Abraham 2010).[14]

To explore some of these possible explanations, we have constructed cost-based and income-based estimates of investment in education for the United States covering the period from 2006 through 2020. Our cost-based estimates incorporate all of the direct spending on education by households, nonprofit institutions serving households, and governments included in the national income and product accounts. To those costs, we add an estimate of the value of the time that students

[13] The cited estimates refer to current-dollar cost-based estimates of the stock of human capital from table B-20 of Kendrick (1976); current-dollar income based estimates from table 8.12 of Jorgenson and Fraumeni (1992b); and estimates of share of the value of the human capital stock accounted for by market income from table 8.16 of Jorgenson and Fraumeni (1992b).

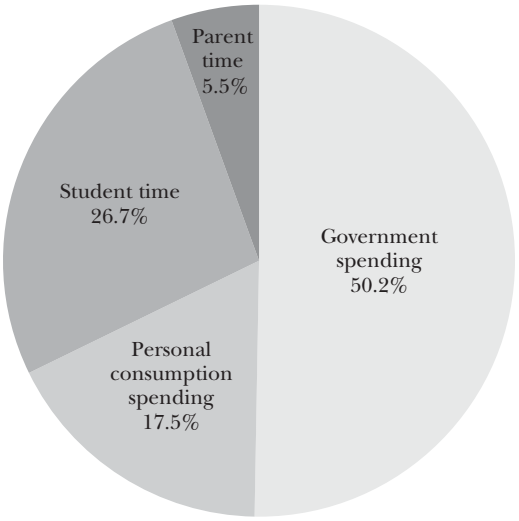
[14] Other reasons to question the income-based estimates include the possibility that earnings reflect factors other than productivity or that relative earnings for different groups of workers might change in the future (Abraham 2010). Even if true, however, it is not clear that either of these would lead to systematic overstatement in the value of investments in education.

age 15 and older devote to education. This estimate is based on school enrollment data from the October education supplement to the Current Population Survey normalized to match enrollment counts reported by the National Center for Education Statistics (NCES) and earnings data from the Annual Social and Economic Supplement (ASEC) to the Current Population Survey. Following the literature, we assume that enrolled students devote 1,300 hours per year to their schooling and value the opportunity cost of that time at the hourly wage of individuals of the same age, sex, and completed education level. As a crude correction for the fact that wages are only a portion of total compensation, we multiply this estimate by 1.235, the average ratio of total compensation to wages and salaries in the national income and product accounts over the 2006–2020 period. Finally, we construct a rough estimate of the value of parent time devoted to their children's schooling using data from the American Time Use Survey on the time adults spend helping children with their schooling. To value this time, we use annual average Current Population Survey data on median weekly earnings for full-time elementary and middle school teachers, converted to an hourly wage assuming that full-time means 40 hours per week and adjusted upwards by a factor of 1.235 to account for components of compensation other than wages and salaries.

Figure 2 shows the relative importance of the different components of the estimated cost of investments in formal education in 2020; the cost shares for other recent years are similar. Government spending accounts for just over half of estimated costs in 2020 (50.2 percent). The next largest contributor is the value of student time at 26.7 percent, with expenditures by households and nonprofit institutions serving households accounting for 17.5 percent and parent time for 5.5 percent.

Our income-based estimates generally follow the approach developed by Jorgenson and Fraumeni. Earnings and hours by age, sex, and education come from the Annual Social and Economic Supplement. Mortality rates by age and sex come from the National Center for Health Statistics' mortality files. As for the cost-based estimates, the information on school enrollments used for the income-based estimates comes from the October education supplement to the Current Population Survey, normalized to match enrollment counts from the National Center for Education Statistics. One difference from the original Jorgenson and Fraumeni calculations is that, using a modification to their approach introduced by Christian (2010), we allow for the possibility that individuals older than age 75 have labor earnings. Another difference is that we use pre-tax wages rather than post-tax wages to estimate the returns to education. As a rough adjustment to account for the value of nonwage compensation, we also multiply the estimated returns to formal education based on the ASEC wages by 1.235. Consistent with the original Jorgenson and Fraumeni work, our baseline estimates assume that the relevant counterfactual for individuals who complete an additional year of schooling is that, had they not done so, their probability of returning to school in the future would have been the same as for a person of the same age but one less year of schooling. Income is assumed to grow at 2 percent annually and the temporal discount factor is set to 4 percent. For this analysis, we count only the market returns to education.

Figure 2
Breakout of the Costs of Investment in Formal Education, 2020



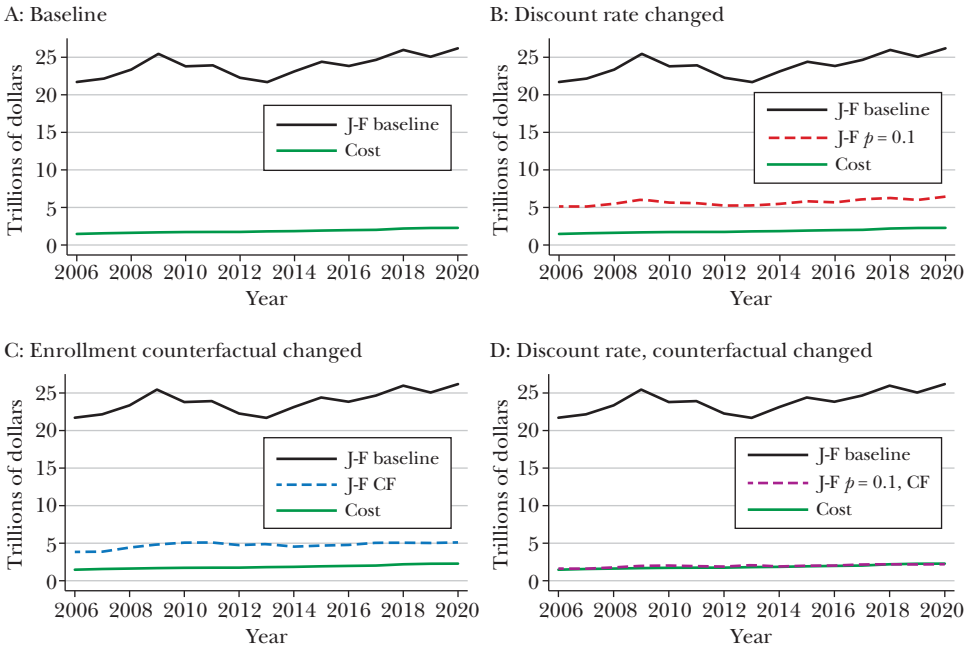
Source: Authors' calculations.
Note: Data on personal consumption and government spending on formal education from the National Income and Product Accounts. Estimated value of student and parent time constructed using data from multiple sources as described in the text.

We would like to know whether varying the assumptions underlying our baseline income-based estimates of investment in education can reconcile the cost-based and income-based estimates. Figure 3 displays the results of this exercise.

Panel A, in the upper left, compares the nominal dollar value of the income-based and cost-based estimates of the value of investment in education under our baseline assumptions. Under these assumptions, the estimated gross market return to investment in education using the income approach is roughly 11 to 15 times the value of the same investment based on the costs of education, depending on the year. These are proportionally larger differences than obtained by comparing the market income-based values reported by Jorgenson and Fraumeni (1992b) to Kendrick's (1976) cost-based estimates for the years 1947 through 1969. The fact that our income-based estimates allow for earnings past age 75, use pre-tax wages rather than post-tax wages, and are adjusted to account for the portion of total compensation that is not wages and salaries makes our ratios larger. The rise in female labor force participation since 1970, and the accompanying shift from nonmarket to market activity (Fraumeni and Christian 2019), also may be a contributing factor. Our numbers are in the same ballpark as the recent estimates for Canada reported by Gu and Wong (2015).

Panel B, in the upper right, modifies our discount rate assumption. As an alternative to assuming an intertemporal discount rate of 4 percent, we assume an intertemporal discount rate of 10 percent. This is admittedly a much higher

Figure 3
Alternative Estimates of US Investment in Human Capital, 2006–2020



Source: Authors' calculations.
Note: J-F Baseline refers to Jorgenson and Fraumeni estimates based on market returns making their baseline assumption about enrollments and assuming a 2 percent earnings growth rate and 4 percent temporal discount rate. J-F $p = 0.10$ changes the assumed temporal discount rate to 10 percent per year. J-F CF changes the assumption about how completing a year of schooling changes future enrollments as described in the text. J-F $p = 0.10$, CF makes both changes. Cost refers to our cost based estimates of investment in education. All figures are in dollars of the indicated year.

discount rate than is typical in the literature, but could perhaps be appropriate for the evaluations made by myopic or risk averse individuals regarding an investment they cannot diversify.[15]

Panel C, in the lower left, modifies our enrollment counterfactual for the income-based estimates concerning the future path of enrollments for someone who misses a year of school. For our alternative counterfactual regarding future education, we assume the same future enrollment probabilities as for someone a year younger with one less year of education, as opposed to the future enrollment probabilities for someone the same age with one less year of education. Because the probability of continuing in school is much higher for people who are on

[15] We do not vary our assumption about the growth rate of future earnings, but lowering the assumed growth rate for earnings by one percentage point would be essentially equivalent to raising the assumed temporal discount rate by one percentage point.

track educationally than for people who have fallen behind, the baseline assumption implies a big difference in expected future earnings for people the same age whose current educational attainment differs by a year. At least in part, however, the differences in future enrollment probabilities are likely to be due to differences in the characteristics of the people who select into staying in school versus dropping out. Our alternative counterfactual assumes, in essence, that someone who fails to complete a year of schooling gets a do-over. Under this alternative counterfactual, the value of completing an extra year of schooling is considerably smaller since it has less effect on a person's future educational attainment.

Panel D, in the bottom right, plots gross investment in education under the income-based approach using both our alternative assumption about the discount rate and our modified counterfactual assumption about probabilities of enrollment for someone who misses a year of school. Changing either our assumption about the discount rate or our assumption about future enrollment probabilities, as is done in panels B and C, reduces the size of the gap, but the income-based estimates are still 2½ to 3½ times as large as the cost-based estimates. Making both changes simultaneously effectively reconciles the average levels of investment estimated using the income-based and cost-based approaches.

Of course, this illustrative set of calculations does not prove that our alternative assumptions are "correct" in any sense. Our calculations do show, however, that methodological assumptions—some fairly obvious like the discount rate, others more subtle like how to model the wage path of those who have fallen a year behind the conventional path in their schooling—can make a large difference in these estimates.

Topics for Future Research

Our discussion has focused mainly on estimates of the value of investments in formal education produced using the cost and income approaches. Estimates produced using these approaches are appealing in that they are conceptually compatible with the existing national income and product accounts. A considerable agenda for future research on the measurement of human capital remains, and here we highlight some major issues.

One issue we have not addressed is the heterogeneity in formal education. Even among those in a given country who have the same number of years of schooling, the value of human capital may depend a great deal on the specific type of schooling a person received. In some contexts, measuring educational attainment by credentials, rather than years of schooling, may be more meaningful. It also may be important to take account of changes related to the characteristics of the students being educated, such as changes in the prevalence of regular versus special education students or students whose native language is not English (Fraumeni et al. 2009). A related complication for cross-country comparisons is that some countries emphasize formal education in preparing young people for

careers, whereas credentials acquired through structured on-the-job training, such as apprenticeship programs, play a larger role in others (Conrad 1992).

Further, the quality of education—what it means, for example, to have a high school diploma or a college degree—may have changed over time. This heterogeneity might be captured by looking either at inputs or at outputs. For example, some of the inputs plausibly affecting the quality of the education a student receives include class size and teacher qualifications such as degrees earned, whether the teacher has been trained in the subject being taught, and years of teaching experience (Fraumeni et al. 2009; UNECE 2016). Alternatively, output measures like test scores may be a useful proxy for the quality of educational attainment, though the skills measured by available tests capture only some of the skills that are likely to affect a person's labor market outcomes (Heckman, Stixrud, and Urzua 2006). Finding adequately reliable and robust ways to account for changes in the quality of education over time and differences in the quality of education across countries would be an important step forward.

Second, there may be a more nuanced way to calculate the nonmarket private returns to education than has been adopted in the literature thus far. In many common tasks of home production, like cleaning the bathroom or doing the laundry, more-educated individuals seem unlikely to enjoy a productivity advantage, but they might have an advantage in others, like engaging with children in ways that enhance their human capital. Finding a way to assess the productivity of more versus less educated individuals in various activities would be a difficult but perhaps not impossible task. One possible approach would be to assign values to time devoted to home production of goods and services that could in principle have been purchased from third-party suppliers.

Finally, the income approach to valuing investment in education treats the returns to education as captured fully by the increment to individual earnings. Although this is a useful starting point, there are almost certainly positive spillovers to others in the population. Positive externalities associated with having a more educated population may include such things as a more informed electorate and a lower crime rate (Abraham and Mackie 2005), as well as the possible agglomeration effects made possible by having larger numbers of highly skilled individuals working together (for example, Puga 2010).

■ *We are especially grateful to Michael Christian for sharing code and data that we have used in constructing income-based estimates of human capital investment. The paper also benefitted from useful conversations with Peter van de Ven, formerly of the OECD, and Ann Lisbet Brathaug of Statistics Norway, and from comments on an earlier draft from Barbara Fraumeni, Erik Hurst, Nina Pavcnik, Timothy Taylor, and Heidi Williams. The views expressed are solely those of the authors and not necessarily those of the US Bureau of Economic Analysis or the US Department of Commerce.*

References

Abraham, Katharine G. 2010. "Accounting for Investments in Formal Education." *Survey of Current Business* 90 (6): 42–53.

Abraham, Katharine G., and Christopher Mackie, eds. 2005. *Beyond the Market: Designing Nonmarket Accounts for the United States*, Washington, DC: National Academies Press.

Abraham, Katharine G., and Justine Mallatt. "Replication data for: Measuring Human Capital." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E172261V1.

Arrow, Kenneth, Partha Dasgupta, Lawrence H. Goulder, Kevin J. Mumford, and Kirsten Oleson. 2012. "Sustainability and the Measurement of Wealth." *Environment and Development Economics* 17 (3): 317–53.

Barro, Robert J., and Jong-Wha Lee. 1993. "International Comparisons of Educational Attainment." *Journal of Monetary Economics* 32 (3): 363–94.

Barro, Robert J., and Jong-Wha Lee. 2001. "International Data on Educational Attainment: Updates and Implications." *Oxford Economic Papers* 53 (3): 541–63.

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

Barro, Robert J. and Jong-Wha Lee. 2021. "Barro-Lee Educational Attainment Dataset." http://www.barrolee.com/ (accessed February 17, 2022)

Barro, Robert J., and Xavier Sala-i-Martin. 1995. *Economic Growth*. New York: McGraw-Hill.

Becker, Gary S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: National Bureau of Economic Research.

Björklund, Anders, and Kjell G. Salvanes. 2011. "Education and Family Background: Mechanisms and Policies." In *Handbook of the Economics of Education,* Volume 3, edited by Eric A. Hanushek, Stephen J. Machin, and Ludger Woessmann, 201–47. Amsterdam: North Holland.

Botev, Jarmila, Balázs Égert, Zuzana Smidova, and David Turner. 2019. "A New Macroeconomic Measure of Human Capital with Strong Empirical Links to Productivity." OECD Working Paper 1575.

Bureau of Economic Analysis. 2021. *NIPA Handbook: Concepts and Methods of the U.S. National Income and Product Accounts*. Washington, DC: US Department of Commerce.

Bureau of Labor Statistics. 1996. "1995 Survey of Employer Provided Training-Employee Results," *US Bureau of Labor Statistics*, December 19. https://www.bls.gov/news.release/sept.nws.htm.

Caplin, Andrew, and John Leahy. 2004. "The Social Discount Rate." *Journal of Political Economy* 112 (6): 1257–68.

Caselli, Francesco, and Antonio Ciccone. 2019. "The Human Capital Stock: A Generalized Approach: Comment." *American Economic Review* 109 (3): 1155–74.

Christian, Michael S. 2009. "Human Capital Accounting in the United States, 1964–2006." Unpublished.

Christian, Michael S. 2010. "Human Capital Accounting in the United States, 1964–2006." *Survey of Current Business* 90 (6): 31–36.

Christian, Michael S. 2014. "Human Capital Accounting in the United States: Context, Measurement, and Application." In *Measuring Economic Sustainability and Progress*, edited by Dale W. Jorgenson, J. Steven Landefeld, and Paul Schreyer, 461–91. Chicago: University of Chicago Press.

Christian, Michael S. 2017. "Net Investment and Stocks of Human Capital in the United States, 1975–2013." *International Productivity Monitor* 33: 128–49.

Cohen, Daniel, and Laura Leker. 2014. "Health and Education: Another Look with the Proper Data." CEPR Discussion Paper 9940.

Cohen, Daniel, and Marcelo Soto. 2007. "Growth and Human Capital: Good Data, Good Results." *Journal of Economic Growth* 12 (1): 51–76.

Conrad, Klaus. 1992. "Comment on D.W. Jorgenson and B.M. Fraumeni, 'Investment in Education and U.S. Economic Growt.'" *Scandinavian Journal of Economics* 94 (Supplement): S71–74.

DeGenova, Mary Kay. 1992. "If You Had Your Life to Live over Again: What Would You Do Differently?" *International Journal of Aging and Human Development* 34 (2): 135–43.

de la Fuente, Angel, and Rafael Doménech. 2015. "Educational Attainment in the OECD, 1960–2010: Updated Series and a Comparison with Other Sources." *Economics of Education Review 48* (1): 56–74.

Diewert, W. Erwin. 2021. *Consumer Price Index Theory*. Unpublished.

Eisner, Robert. 1978. "Total Incomes in the United States, 1959 and 1969." *Review of Income and Wealth* 24 (1): 41–70.

Eisner, Robert. 1985. "The Total Incomes System of Accounts." *Survey of Current Business* 65 (1): 24–48.

Eisner, Robert. 1989. *The Total Incomes System of Accounts*. Chicago: University of Chicago Press.

European Commission, International Monetary Fund, Organisation for Economic Cooperation and Development, United Nations, and World Bank. 2009. *System of National Accounts 2008*. New York: United Nations.

Eurostat. 2022. "Adult Learning Statistics." https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Adult_learning_statistics (accessed February 19, 2022).

Fraumeni, Barbara M., and Michael S. Christian. 2019. "Accumulation of Human Capital in the United States, 1975–2012: An Analysis by Gender." NBER Working Paper 25864.

Fraumeni, Barbara M., Marshall B. Reinsdorf, Brooks B. Robinson, and Matthew P. Williams. 2009. "Price and Real Output Measures for the Education Function of Government: Exploratory Estimates for Primary and Secondary Education." In *Price Index Concepts and Measures,* edited by W. Erwin Diewert, John Greenlees, and Charles R. Hulten, 373–403. Chicago: University of Chicago Press.

Gong, Yifan, Lance Lochner, Ralph Stinebrickner, and Todd R. Stinebrickner. 2021. "The Consumption Value of College." NBER Working Paper 26335.

Goujon, Anne, et al. 2016. "A Harmonized Dataset on Global Educational Attainment between 1970 and 2060—An Analytical Window into Recent Trends and Future Prospects in Human Capital Development." *Journal of Demographic Economics* 82 (3): 315–63.

Gu, Wulong, and Ambrose Wong. 2015. "Productivity and Economic Output of the Education Sector." *Journal of Productivity Analysis* 43 (2): 165–82.

Hall, Robert E., and Charles I. Jones. 1999. "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" *Quarterly Journal of Economics* 114 (1): 83–116.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411–82.

Jones, Richard, and Valerie Fender. 2011. "Human Capital Estimates, 2010." Office for National Statistics. https://webarchive.nationalarchives.gov.uk/ukgwa/20160106223656mp_/http://www.ons.gov.uk/ons/dcp171766_248886.pdf.

Jorgenson, Dale W., and Barbara M. Fraumeni. 1989. "The Accumulation of Human and Nonhuman Capital, 1948–84." In *The Measurement of Saving, Investment, and Wealth*, edited by Robert E. Lipsey and Helen Stone Tice, 227–86. Chicago: University of Chicago Press.

Jorgenson, Dale W., and Barbara M. Fraumeni. 1992a. "Investment in Education and U.S. Economic Growth." *Scandinavian Journal of Economics* 94 (Supplement): S51–70.

Jorgenson, Dale W., and Barbara M. Fraumeni. 1992b. "The Output of the Education Sector." In *Output Measurement in the Service Sectors*, edited by Zvi Griliches, 303–41. Chicago: University of Chicago Press.

Katz, Arnold J. 2015. "A Primer on the Measurement of Net Stocks, Depreciation, Capital Services, and Their Integration." Bureau of Economic Analysis Working Paper 2015-6.

Kendrick, John W. 1976. *The Formation and Stocks of Total Capital*. New York: National Bureau of Economic Research.

Klenow, Peter. J., and Andrés Rodríguez-Clare. 1997. "The Neoclassical Revival in Growth Economics: Has It Gone Too Far?" In *NBER Macroeconomics Annual 1997*, Volume 12, edited by Ben Bernanke and Julio J. Rotemberg, 73–114. Cambridge, MA: MIT Press.

Kraay, Aart. 2019. "The World Bank Human Capital Index: A Guide." *World Bank Research Observer* 34 (1): 1–33.

Kuznets, Simon. 1961. *Capital in the American Economy: Its Formation and Financing*. Princeton NJ: Princeton University Press.

Lange, Glenn-Marie, Quentin Wodon, and Kevin Carey, eds. 2018. *The Changing Wealth of Nations: Building a Sustainable Future*. Washington, DC: World Bank Group.

Le, Trinh, John Gibson, and Les Oxley. 2005. "Measures of Human Capital: A Review of the Literature." New Zealand Treasury Working Paper 05/10.

Levesque, Christopher, and Jeanne Batalova. 2022. "Frequently Requested Statistics on Immigrants and Immigration in the United States," *Migration Policy Institute,* February 17. https://www.migrationpolicy.org/article/frequently-requested-statistics-immigrants-and-immigration-united-states.

Liu, Gang. 2011. "Measuring the Stock of Human Capital for Comparative Analysis: An Application of the Lifetime Income Approach to Selected Countries." OECD Statistics Working Paper 2011/06.

Liu, Gang. 2014. "Measuring the Stock of Human Capital for International and Intertemporal Comparisons." In *Measuring Economic Sustainability and Progress*, edited by Dale W. Jorgenson, J. Steven Landefeld, and Paul Schreyer, 493–544. Chicago: University of Chicago Press.

Liu, Gang, and Barbara Fraumeni. 2020. "A Brief Introduction to Human Capital Measures." NBER Working Paper 27561.

Lucas, Robert E., Jr. 1988. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22 (1): 3–42.

Managi, Shunsuke, and Pushpam Kumar, eds. 2018. *Inclusive Wealth Report 2018: Measuring Sustainability and Well-Being*. Nairobi, Kenya: United Nations Environment Programme.

O'Mahony, Mary, and Philip Stevens. 2009. "Output and Productivity Growth in the Education Sector: Comparisons for the US and UK." *Journal of Productivity Analysis* 31 (3): 177–94.

Puga, Diego. 2010. "The Magnitude and Causes of Agglomeration Economies." *Journal of Regional Science* 50 (1): 203–19.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5 Part 2): S71–102.

Samans, Richard, Saadia Zahidi, Till Alexander Leopold, and Vesselina Ratcheva. 2017. *The Global Human Capital Report 2017: Preparing People for the Future of Work*. Geneva, Switzerland: World Economic Forum.

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

United Nations Economic Commission for Europe (UNECE). 2016. *Guide on Measuring Human Capital*. Geneva, Switzerland: United Nations.

United Nations Economic Commission for Europe (UNECE). 2020. *Satellite Account for Education and Training: Compilation Guide*. Geneva, Switzerland: United Nations.

World Bank. 2020. *The Human Capital Index 2020 Update: Human Capital in the Time of COVID-19*. Washington, DC: World Bank Group.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 131–156

Expected and Realized Inflation in Historical Perspective

Carola Binder and Rupal Kamdar

Economists have long discussed the importance of inflation expectations for economic outcomes. An early and prominent example is the "Fisher effect"—Irving Fisher's famous hypothesis that expected inflation is equal to the difference between nominal and real interest rates (Fisher 1896, 1907, 1930). Fisher did not claim that this insight was original, but referred to antecedents including a 1740 pamphlet by William Douglass as well as better-known earlier writers like John Stuart Mill and Alfred Marshall (Dimand 1999).

The basic Fisher effect relationship raises theoretical and empirical questions about inflation expectations. How are inflation expectations determined: namely, are they backward-looking or forward-looking? Additionally, is there a two-way relationship between inflation expectations and inflation itself? For example, past inflation may shape current inflation expectations, but current inflation expectations may also shape current and future inflation. A variety of mechanisms imply that inflation and inflation expectations are interconnected. For instance, if workers expect high inflation, they may ask their employer for a raise so that their incomes keep up with the cost of living. However, the employer may then raise prices to compensate for the higher wage bill, creating inflation. In this example, higher inflation expectations create inflation. Similarly, suppose a firm changes its prices only once a quarter. Then, the firm will consider how much its costs will rise over the course of the upcoming quarter when setting its price—again, higher inflation

■ *Carola Binder is Associate Professor of Economics, Haverford College, Haverford, Pennsylvania. Rupal Kamdar is Assistant Professor of Economics, Indiana University, Bloomington, Indiana. Their email addresses are cbinder1@haverford.edu and rkamdar@iu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.131.

expectations are self-fulfilling and result in higher inflation. On the other hand, if inflation has been high in recent months, individuals who extrapolate their beliefs from recent inflation may increase their inflation expectations—that is, higher inflation may result in higher inflation expectations.

These questions and mechanisms are at the heart of monetary policy today. If inflation expectations were both accurate and adjusted in real time precisely with nominal interest rates, then it would be difficult for the Federal Reserve to alter real interest rates. Therefore, the conduct of modern monetary policy relies on an understanding of inflation expectations. Indeed, Federal Reserve Chair Jerome Powell (2019) remarked in Congressional testimony that "in our thinking, inflation expectations are the most important driver in actual inflation." Similarly, a staff report of the Federal Reserve Bank of New York explains that "monitoring and managing consumers' expectations have become primary goals of policy makers, and are central components of modern monetary policy" (Armantier et al. 2016).

In this paper, we provide historical context for the relationship between realized and expected inflation—a relationship that, even now, provokes considerable controversy (Rudd 2021). We begin with a discussion of early theories about how inflation expectations are formed. Next, we discuss measures of inflation expectations and their empirical relationship to current and future realized inflation. Lastly, we provide a narrative account of the relationship between expected and realized inflation in the United States during key periods, including the Great Depression of the 1930s, the Great Inflation of the 1970s, the Great Recession of 2008–2009, and the recent COVID-19 pandemic.

How Are Inflation Expectations Set?

Once Irving Fisher popularized the importance of inflation expectations for economic outcomes, it became necessary to model inflation expectations to test hypotheses such as the Fisher effect. Initially, Fisher modeled expectations as a weighted sum of current and past inflation. This backward-looking approach is known as adaptive expectations. However, in the late 1960s and early 1970s, other economists noted shortcomings in the slow-to-adjust expectations and developed rational expectations. Under rational expectations, a forward-looking framework, economic agents use all relevant information to determine their expectations.

"To Be Forewarned Is to Be Forearmed"

Irving Fisher (1867–1947) is well-known for his role in the development and popularization of price indexes,[1] as well as his contributions to economic thought on interest rates, inflation, and expected inflation. In reference to the difference between expected and realized inflation, Fisher (1911) argued that "the real evils

[1] For a discussion in this journal of the development of price indexes preceding and contemporaneous with Fisher, see Persky (1998).

of changing price levels do not lie in these changes per se, but in the fact that they usually take us unawares."[2] Put another way, individuals are unable to forecast inflation perfectly, so they cannot be sure how far money will go in terms of buying goods in the future. Thus, any contracts written in terms of money will result in an uncertain payoff in terms of goods. Fisher believed that improvements in knowledge could reduce the difference between expected and realized inflation and therefore mitigate the cost of inflation:

> It has been shown that to be forewarned is to be forearmed, and that a foreknown change in price levels might be so taken into account in the rate of interest as to neutralize its evils. While we cannot expect our knowledge of the future ever to become so perfect as to reach this ideal, . . . nevertheless every increase in our knowledge carries us a little nearer that remote ideal.

However, Fisher (1911) also acknowledged that cognitive constraints and inattention would make it difficult for the average businessman, with limited theoretical knowledge, to form reasonably accurate inflation expectations. Fisher wrote this during the classical gold standard era, when prices were relatively stable over the longer run (Klein 1975). Indeed, an "implication of the tendency for price levels to revert toward a long-run stable value under the gold standard was that it insured a measure of predictability with respect to the value of money: though prices would rise or fall for a few years, inflation or deflation would not persist," as noted in Bordo (1981, p. 11). The difficulty of forming accurate inflation expectations would only be exacerbated by the tumultuous times to come.

As World War I began in 1914, many countries suspended the convertibility of currency to gold. Ultimately, this brought about the end of the classical gold standard and resulted in a large increase in inflation, reaching as high as 20 percent in 1917. With the deep recession of 1920–1921 came an extreme deflation, with prices briefly falling at a rate of 20 percent a year.[3] Even as the world returned to gold in 1925, this time under the gold exchange standard,[4] prices were less stable than under the classical gold standard (Bordo 1981). The volatility of inflation resulted

[2]This idea was also not altogether new. Lowe (1823, p. 96), for example, argued that "contracts for a series of years ought to be made with reference to the power of money in purchasing the necessaries and comforts of life."

[3]Estimates of inflation and deflation are calculated as the percent change from a year ago of the general price level. The data for the general price level were obtained from the NBER Macrohistory Database (FRED series M04051USM324NNBR). Note that inflation as calculated using the wholesale price level (FRED series M04049USM052NNBR) is more volatile, with prices rising by as much as 40 percent in 1917 and falling by as much as 50 percent in 1921.

[4]Under the classic gold standard, participating countries guaranteed their currency was convertible to a specified amount of gold, and thus needed to hold gold reserves. Under the gold exchange standard, participating countries guaranteed their currency could be exchanged for either gold directly at a fixed rate or to another currency which could then be converted to gold at a fixed rate. The United Kingdom and the United States only held gold reserves. Other countries on the gold exchange standard could hold gold, dollars, or pounds as reserves. The gold exchange standard ended in 1931 when Britain withdrew.

in increased attention to prices and inflation. In response to public demand for information about price indices, Fisher established the Index Number Institute in the 1920s, which for some years operated out of his home and sold information to newspapers. By 1929, his wholesale price index reached a newspaper audience of 5 million.[5]

The Great Depression began in 1929 and brought with it a sharp deflation. Fisher believed that the deflation was unanticipated and was thus costly (Allen 1977). Accordingly, he was interested in understanding how people formed their inflation expectations. These expectations played a central role not only in his understanding of the Great Depression but also in his *Theory of Interest* (1930). In Part IV, Chapter XIX of that book, Fisher wrote:

> How is it possible for a borrower or lender to foresee variations in the general price level with the resultant increase or decrease in the buying power of his money? A change in the value of money is hard to determine. Few business men have any clear ideas about it . . . Yet it may be true that they do take account, to some extent at least, even if unconsciously, of a change in the buying power of money . . . If inflation is going on, they will scent rising prices ahead . . . And today especially, foresight is clearer and more prevalent than ever before. The business man makes a definite effort to look ahead not only as to his own particular business but as to general business conditions, including the trend of prices.

Adaptive Expectations

Understanding how inflation expectations are formed and how to model them was key to Fisher's research agenda. For example, the "Fisher effect" hypothesis that the nominal interest rate is the sum of the real interest rate and expected inflation predicts a strong, positive correlation between nominal interest rates and expected inflation. To test this hypothesis, Fisher modeled expected inflation as a weighted average of current and past inflation, reasoning that "price changes do not exhaust their effects in a single year but manifest their influence with diminishing intensity." Fisher's "adaptive" approach to modeling expectations had the benefits of simplicity and feasibility.

However, there are obvious shortcomings as well: for example, it assumes that the structure that generates inflation expectations stays the same over time. For the United States, Fisher found the highest correlation between nominal interest rates and a weighted average of past inflation over 20 years—a time lag later critiqued by Cagan (1965) as implausibly long.

While Fisher's analysis emphasized how realized inflation might affect expected inflation, it did not explore the reverse direction of causality. The *bidirectional*

[5]Vogt, Arthur. 2020. "Fisher, Irving." In *Encyclopedia of Mathematics*. https://encyclopediaofmath.org/wiki/Fisher,_Irving (accessed May 20, 2022).

relationship between realized and expected inflation was made prominent by Edmund Phelps (1967) and Milton Friedman (1968). Before their work, it was widely believed that policymakers could always use expansionary monetary policy to reduce unemployment at the cost of more inflation. This trade-off, known as the Phillips curve, was believed to hold even in the long run (Gordon 2018, as discussed in this journal by Hall and Sargent 2018). But Friedman (1968, p. 11) claimed that the trade-off between unemployment and inflation was actually temporary because it "comes not from inflation per se, but from unanticipated inflation"—that is, the difference between realized and expected inflation. Friedman was arguing that expected inflation itself is a determinant of unemployment and therefore inflation.

Furthermore, Friedman (1968, p. 11) continued by commenting that "unanticipated inflation . . . generally means . . . a rising rate of inflation." This is the direct result of the assumption, shared by Phelps, that inflation expectations were formed in an adaptive or backward-looking manner. This view implied that policymakers could only keep unemployment below its "natural rate"[6] in the short run by "accelerating" inflation to stay ahead of the public's backward-looking expectations. In other words, inflation must be higher than what was expected to reduce unemployment below its natural rate. But if inflation expectations are formed in a backward-looking manner, then expected inflation for the next period will rise. In order to maintain the low unemployment rate, inflation must once again surpass the newly-revised expectations, and so on.

However, modeling expectations as backward-looking is unlikely to be appropriate in all contexts. In fact, Friedman acknowledged this shortcoming; for instance, he argued that in settings with high inflation, the adjustment of expectations would likely occur rapidly.

Rational Expectations

These concerns led Lucas (1972, 1973) and Sargent and Wallace (1976) to modify the models of Friedman and Phelps by incorporating the assumption of rational expectations (attributed to Muth 1961). If inflation expectations are rational, they incorporate all information that is useful in forecasting future inflation. For example, Sargent (1973, p. 447) showed that interest rates contain information that is useful in predicting inflation, which "implies that it is probably inadequate to hypothesize that expectations of inflation are simply naive extrapolations of past rates of inflation, since that involves supposing that readily available information about the subsequent course of inflation goes unused."

Lucas's work based on rational expectations found that unanticipated changes in inflation were required to change output or unemployment, much like Friedman and Phelps argued. That is, expectations affect output, unemployment, and, thus, inflation. The rational expectations revolution that followed these innovations was described as "one of the defining features in the rebuilding of macroeconomics" by

[6]The natural rate of unemployment is the lowest level of unemployment an economy can sustain without rising inflation.

Coibion and Gorodnichenko (2012). Indeed, rational expectations are used in the New Keynesian approach, which started in the late 1970s and 1980s and has become the dominant modeling approach for macroeconomics (as discussed in this journal by Galí 2018).

The New Keynesian model features firms seeking to maximize their profits while subject to a pricing friction. A common friction is Calvo (1983) pricing, which assumes only some firms will be allowed to change their prices each period. This friction results in forward-looking pricing decisions. Firms set their price not only for the current period, but possibly for several future periods in which they will not be allowed to change their price. Overall, firms' inflation expectations influence the prices that firms set, and thus, inflation expectations affect realized inflation.

Inflation in the New Keynesian model is summarized by the famed New Keynesian Phillips curve, which says the inflation rate is determined by two factors: expected inflation by firms and the output gap (Galí 2018). The second term captures how an economy is operating relative to full employment and potential GDP, and reflects the possibility of a tradeoff between unemployment and inflation, reminiscent of the original Phillips curve. However, in contrast to the original Phillips curve, the New Keynesian Phillips curve features expected inflation as a determinant of inflation.

Are Measures of Inflation Expectations Related to Realized Inflation?

Given the theoretical relationships between inflation and inflation expectations, it is natural to ask if this relationship holds empirically. To answer this question, direct measures of inflation expectations are necessary and can be derived from surveys or asset prices. These measures of inflation expectations are indeed strongly correlated with contemporaneous and future inflation when calculated over long samples, such as decades. However, over shorter time frames, such as a few years, the relationship can weaken.

Survey Measures of Expected Inflation

Over the years, many surveys have been created to solicit direct measures of inflation expectations of professional forecasters, households, and firms. The questions posed to respondents have varied along two key dimensions. First, questions have differed in which price index they refer; for example, surveys of households tend to ask for inflation expectations for "prices in general," while professional forecaster surveys tend to specify the price index, such as the Consumer Price Index. Second, questions have differed across the horizon of inflation expectations requested; for example, some questions ask about short horizons, such as the coming quarter, year, or two years, while others ask about long horizons, such as the next five or ten years.

The oldest, continuous survey of inflation expectations is the Livingston Survey. It was started in 1946 by financial journalist Joseph Livingston for the *Philadelphia*

Inquirer. Respondents, who were economists working in industry, government, banking, and academia, were asked twice a year to provide their forecasts for over a dozen variables, including the Consumer Price Index, for several time horizons. Initially, the effort was for the purposes of journalism rather than academic research (as discussed in this journal by Thomas 1999). However, economists in the 1970s went in search of expectations data to test the new theory of rational expectations and found the Livingston Survey. Given the new interest in the survey, Livingston partnered with the Philadelphia Federal Reserve to manage and share the data with economists in a centralized manner. When Livingston died in 1989, the Philadelphia Federal Reserve took the survey over (Croushore 1997). The Livingston survey is still used today, but less often than newer surveys available at higher frequencies.[7]

In 1969, the National Bureau of Economic Research and the American Statistical Association partnered to develop a new quarterly survey. The purpose was to create a representative survey of professional forecasters with sufficient frequency and a long, consistent time series (Zarnowitz and Braun 1993). Today the survey is known as the Survey of Professional Forecasters, includes approximately 40 forecasters, and is conducted by the Federal Reserve Bank of Philadelphia.[8] At first, the only inflation forecasts collected were for inflation as measured by the gross domestic product deflator, but additional forecast variables and horizons have been introduced over the years. Forecasts of Consumer Price Index inflation are available since 1981:III for shorter time horizons, and since 1991:IV for the ten-year horizon.

Beyond surveys of professional forecasters, there are also surveys of households. Notably, the Michigan Survey of Consumers was created by George Katona at the University of Michigan's Survey Research Center in 1946. Katona believed consumers were powerful economic agents whose consumption and savings decisions could induce expansions and recessions. He further posited that consumption and savings choices are affected by expectations, and thus he set out to measure economic expectations (Curtin 2016). Before 1959, the Michigan Survey of Consumers was conducted irregularly—sometimes twice, sometimes three times annually. The survey was then quarterly from 1959 through the end of 1977 and has been monthly since 1978. One-year-ahead inflation expectations of consumers have been solicited monthly since 1978, and five- to ten-year-ahead expectations have been collected irregularly since 1979 and monthly since 1990. Today, the survey consists of roughly 500 consumers each month and some respondents are surveyed twice.[9]

When referring to survey data on inflation expectations in this essay, we focus on the Livingston Survey, the Survey of Professional Forecasters, and the Michigan Survey of Consumers. However, we should mention some newer surveys. The Federal Reserve Bank of New York started its own Survey of Consumer Expectations

[7]The Livingston survey is available at https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/livingston-survey.
[8]The data is available at https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters.
[9]For more information and data from this survey, see https://data.sca.isr.umich.edu/.

in 2013. This monthly survey of roughly 1,300 consumers solicits expectations on topics such as inflation, job prospects, and earnings, and sometimes it is supplemented by modules on special topics. Despite its short time series, a strength of this survey is that respondents are surveyed monthly for up to twelve consecutive months, which allows for analysis of how a given individual's beliefs change across time.[10]

The Federal Reserve Bank of Atlanta started the Business Inflation Expectations Survey in 2011. The monthly survey includes about 300 panelists representing businesses of a range of sizes that are headquartered in the southeastern states within the district of the Atlanta Federal Reserve.[11] One implication of the New Keynesian Phillips curve is that the inflation expectations that matter for realized inflation are those of firms. However, in the United States, a long-running, nationally representative survey of inflation expectations by firms is not available. That said, recent evidence suggests consumer expectations may be a good proxy for firm expectations. For instance, Coibion, Gorodnichenko, and Kumar (2018) show that in New Zealand consumer expectations are similar to firm expectations.

Market Measures of Inflation Expectations

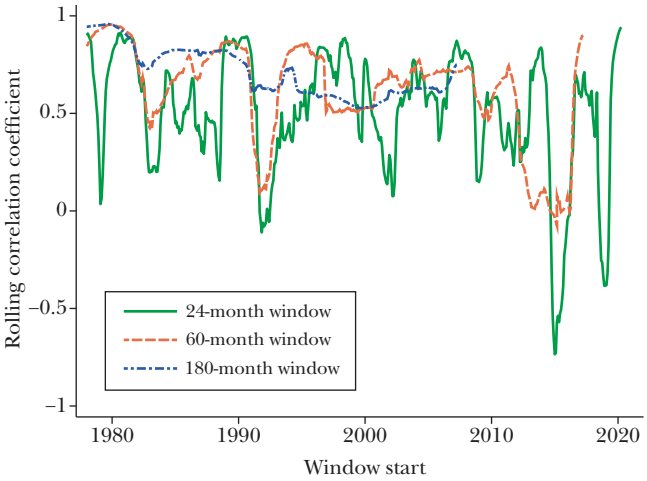
Inflation expectations can also be derived from financial markets. The main market-based measure of inflation expectations uses Treasury Inflation-Protected Securities (TIPS). This approach provides high-frequency measures of inflation expectations but also conflates inflation expectations with other risks, and it has a shorter time series than some of the aforementioned surveys.

When the federal government uses Treasury Inflation-Protected Securities to borrow, the principal is adjusted based on Consumer Price Index inflation. If there is inflation, the principal rises; if there is deflation, the principal falls. Furthermore, the interest rate paid on TIPS applies to the adjusted principal. Thus, one can compare what investors are willing to pay for a Treasury security that is not inflation-protected to an otherwise identical Treasury security that is inflation-protected and calculate what is referred to as the "breakeven" point. This point reveals the inflation compensation that market participants require to avoid inflation exposure. Gürkaynak, Sack, and Wright (2010) estimate a nominal Treasury yield curve as well as a TIPS yield curve, from which they compute "breakeven inflation," for any time horizon. These inflation compensation measures are largely driven by inflation expectations but are also affected by other factors. For instance, the market for TIPS is smaller and less liquid than the market for nominal Treasuries, so some of the differences in yields is driven by differential liquidity premia. Furthermore, the

[10] The Survey of Consumer Expectations is available at https://www.newyorkfed.org/research/staff_reports/sr800.html.

[11] The Business Inflation Expectations Survey is available at https://www.atlantafed.org/research/inflationproject/bie.

Figure 1
Rolling Window Correlation of Inflation and Expected Inflation



Source: Binder and Kamdar (2022).
Note: Correlations between the median, Michigan Survey of Consumers, one-year horizon inflation expectations, and Consumer Price Index inflation are plotted for rolling windows with lengths of 24, 60, and 180 months. Correlations are plotted at the start of the sample window.

difference in yields also reflects inflation risk premia—investors recognize that realized inflation may differ from their expected inflation.

Another shortcoming is that Treasury Inflation-Protected Securities were not traded until 1997. To create a longer time series, researchers have calculated inflation expectations based on "synthetic" TIPS. For example, Groen and Middeldorp (2013) use the relationship between TIPS yields and a selection of 108 long-running time series—including nominal yields on Treasury securities, measures of economic growth, measures of financial stress like volatility and credit spreads, private-sector expectations of the GDP deflator from the Survey of Professional Forecasters, and the "output gap" between the actual and natural rate of unemployment—to construct "synthetic TIPS" rates since 1971.

Correlations between Realized and Expected Inflation

Inflation and measures of expected inflation are highly correlated over long periods of time. For example, from January 1978 to February 2022, the correlation between one-year horizon inflation expectations from the Michigan Survey of Consumers and Consumer Price Index inflation is 0.92. For any 15-year window, the correlation is at least 0.52 and averages 0.71. However, over shorter windows, the correlation is occasionally near zero and at times even negative. Figure 1 plots correlations between one-year horizon inflation expectations from the Michigan Survey of Consumers and Consumer Price Index inflation for rolling windows with

Table 1
Correlations between Inflation, Future Inflation, and Expected Inflation

	Inflation	*Future inflation*	*Michigan survey*
Panel A. Correlation coefficients 1981–2021			
Inflation	1.00		
Future inflation	0.29	1.00	
Michigan survey of consumers	0.74	0.19	1.00
Survey of professional forecasters	0.80	0.46	0.65
Panel B. Correlation coefficients 2011–2021			
Inflation	1.00		
Future inflation	–0.08	1.00	
Michigan survey of consumers	0.36	0.03	1.00
Survey of professional forecasters	0.35	–0.03	–0.31

Source: Binder and Kamdar (2022).
Note: Correlations between Consumer Price Index inflation, one-year-ahead Consumer Price Index Inflation, Michigan Survey of Consumers one-year inflation expectations, and Survey of Professional Forecasters one-year Consumer Price Index inflation expectations are computed at quarterly frequency from 1981:III to 2021:I (159 observations) and from 2011:I to 2021:I (41 observations).

lengths of 2, 5, and 15 years. Overall, the correlation between consumer inflation expectations and realized inflation is strong and largely stable over long samples; however, there are frequent deviations over shorter samples.

This pattern is also seen in the correlation amongst different measures of inflation expectations, as well as the correlations with future inflation. That is, over long samples there are strong, positive correlations; but over shorter samples, these series can be uncorrelated or negatively correlated. Table 1 reports the pairwise correlation coefficients between current inflation, the Michigan Survey of Consumer's one-year-ahead inflation expectations, the Survey of Professional Forecaster's one-year-ahead inflation expectations, and next year's realized inflation. Panel A shows that over the long, four-decade window of 1981 to 2021, all measures are positively correlated. For example, both consumer and professional inflation expectations have a strong, positive correlation with current inflation (coefficients of 0.74 and 0.80, respectively). Furthermore, consumer and professional inflation expectations have a positive correlation with future inflation (albeit to a smaller extent, with coefficients of 0.19 and 0.49, respectively). The measures of inflation expectations of consumers and professionals are also highly correlated with each other over this long sample, with a correlation coefficient of 0.65.

Over the shorter one-decade window from 2011 to 2021 shown in panel B, consumer and professional inflation expectations are still positively correlated with current inflation (coefficients of 0.36 and 0.35, respectively). However, the rest of the correlations in panel B are close to zero or even negative. For instance, consumer and professional expectations are uncorrelated with future inflation (coefficients of 0.03 and -0.03, respectively). Furthermore, the measures of inflation expectations are negatively correlated with each other. Overall, there are strong,

positive correlations over long samples but breakdowns in the correlations over short samples. This suggests that there is much to be learned about the nature and stability of the relationship between inflation and the inflation expectations of various economic agents.

Although the correlations in Table 1 between inflation expectations and future inflation may appear low, note that inflation is difficult to forecast, and inflation expectations are indeed one of the best ways to predict inflation. Ang, Bekaert, and Wei (2007) show that in forecasting future inflation, survey-based measures of expectations outperform a variety of more complicated econometric models such as time series models, Phillips curve-inspired models, and term structure models. Surveys of professionals such as the Livingston Survey and the Survey of Professional Forecasters produce the most accurate forecasts, but the accuracy of inflation forecasts from the Michigan Survey of Consumers is not far behind.

Modeling the Relationship between Expected and Realized Inflation

Researchers have sought to go beyond these correlations and model the extent to which shifts in inflation expectations cause changes in future inflation. The identification difficulties are formidable and potential solutions involve both data and modeling.

One approach is to estimate the New Keynesian Phillips curve. Estimates of this relationship based on survey data of inflation expectations can be sensitive to the choice of survey, sample, and inflation series, and it is not clear which survey expectations to use given that surveys of firms' inflation expectations are limited. However, Coibion and Gorodnichenko (2015), Coibion, Gorodnichenko, and Kamdar (2018), and Coibion, Gorodnichenko, and Kumar (2018) argue in favor of using consumer inflation expectations as a proxy for the expectations of price-setters in firms. Doing so results in stable estimates of the New Keynesian Phillips curve that imply that inflation responds strongly to changes in short-run expected inflation.

Substantial debates have surrounded whether the New Keynesian Phillips curve can represent inflation dynamics in a realistic manner (Cogley and Sbordone 2008). An alternative approach is to estimate a vector autoregression—that is, an essentially model-free approach that only uses past values of macroeconomic variables to predict future values. A summary of the work on inflation expectations using vector autoregressions is that shocks to expectations—especially longer-run expectations—do affect realized inflation and the effect is persistent. Conversely, shocks to actual inflation do not significantly affect long-run or short-run inflation expectations (presumably because such shocks are expected to be temporary). Moreover, long-run expectations significantly affect short-run expectations, but not vice versa (Clark and Nakata 2008; Clark and Davig 2008; Clark and Davig 2009).

While estimates of the New Keynesian Phillips curve and vector autoregression models help understand the relationship between inflation and inflation expectations over long samples, these approaches struggle to capture the subtleties of the relationship between inflation and inflation expectations during short-lived but major events.

A Narrative History of Inflation Expectations and Inflation

A narrative account of inflation and expected inflation in the United States offers compelling evidence of the importance of long-run expectations and policy regime changes in inflation dynamics, as well as a more nuanced interpretation of the relationship between actual and expected inflation.

Inflation Expectations in the Great Depression

Although inflation expectations surveys are not available for the Great Depression era, a cottage industry of academic research beginning in the early 1990s has attempted to pinpoint whether and when the deflation of 1930–1932 was anticipated and when consumers began to expect a return to positive inflation (Romer and Romer 2013, p. 68). Binder (2016) has categorized the approaches in this literature as time-series approaches (Cecchetti 1992; Dorval and Smith 2013), market approaches based on asset prices (Hamilton 1992), and narrative approaches (Nelson 1991; Romer and Romer 2013).

This literature largely finds that the deflation at the start of the Great Depression was unanticipated. For example, Hamilton (1992) shows that futures prices were above spot prices for most commodities, indicating that investors did not expect prices to fall. This finding is consistent with Fisher's (1933) "debt deflation theory," in which unanticipated deflation results in unexpectedly high real interest rates and constitutes a transfer of real wealth from debtors to creditors. This increase in borrowers' indebtedness causes financial distress, including bankruptcies and impaired credit intermediation (Bernanke 1983). The resulting credit contraction, in turn, reduces aggregate demand and leads to further deflation, in a "vicious spiral" (Fisher 1933, p. 346).

After the start of the Great Depression and once deflation had set in, to the extent inflation expectations were backward-looking, inflation expectations would have decreased substantially and likely turned negative. Thus, even as nominal interest rates were low, real interest rates would have been high, dampening demand, thus deepening the Depression and placing additional downward pressure on prices (Schwartz 1981). The restoration of positive inflation expectations was crucial for ending the deflationary spiral and enabling the recovery. President Franklin D. Roosevelt's new macroeconomic policy regime, which began in March 1933 and included the exit from the gold standard, successfully shifted inflation expectations sharply upward in a forward-looking manner. Lacking direct survey- or market-based measures of inflation expectations during this era, the literature has relied on a variety of other approaches but has consistently reached the same conclusion: inflation expectations rose rapidly (Eggertsson 2008). For example, Jalil and Rua (2016) document the rise in the frequency of inflation discussions in the news as a proxy for higher inflation expectations, while Temin and Wigmore (1990) provide anecdotal evidence in line with a rise in inflation expectations, such as a large shift of assets from cash, which loses value during inflation, to the stock market, which tends to rise with inflation.

Following the rise in inflation expectations, a rapid economic expansion ensured. Inflation itself also rose quickly, likely in part because of the rise in inflation expectations. The recovery from the Great Depression lasted from 1933 through 1937, when the Fed raised interest rates in fear of rising inflation. This dramatic episode demonstrates how expectations of deflation or inflation can be self-fulfilling and illustrates the power of a regime change to shift beliefs about inflation in a sudden and drastic manner.

Inflation Expectations from World War II through the Korean War

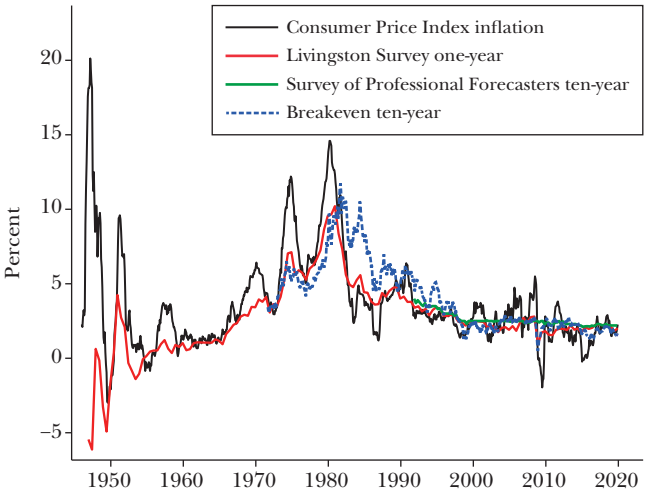
Major wars are often associated with large swings in both actual and expected inflation, and this was certainly the case in World War II. When the United States entered the war, the Federal Reserve issued a statement that it was "prepared to use its powers to assure at all times an ample supply of funds for financing the war effort . . ." (Board of Governors 1943, p. 1). In the wartime environment of fiscal dominance, inflation was volatile and inflation expectations were likely unanchored. To combat wartime inflation without raising interest rates, the government, supported by the Fed, imposed price and wage controls, rationing, and tighter consumer credit regulations. In 1946, a burst of inflation then followed the removal of price controls (Rockoff 1981). According to Friedman and Schwartz (1963, p. 597), "in the immediate postwar years, the public at large anticipated a substantial decline in prices at some future date. The mildness of the 1948–1949 recession and the failure of prices to retreat more than slightly from their postwar highs must have weakened that expectation, and the outbreak of the Korean War gave it the coup de grace."

With the start of the Korean War in 1950, consumer inflation expectations rose, driven in part by backward-looking memories of inflation and scarcities associated with World War II (Binder and Brunet 2022).[12] David Ginsburg (1952, p. 518), a contemporary observer, wrote that inflation was "mostly speculative . . . consumers manifested in the market their anticipations of future shortages and price increases—and thus, in large measure, brought about with their fears the very conditions against which they sought to insure themselves." That is, expectations of inflation led to purchasing of goods in ways which resulted in shortages—which in turn led to actual inflation.

The Livingston Survey of inflation expectations had just begun at this time, and the survey evidence is consistent with this story. Figure 2 plots the Livingston Survey's median, one-year-ahead inflation expectations along with realized inflation and some more recent measures of inflation expectations. Notice that the median Livingston inflation expectation was approximately -5 percent in late 1946, as people expected the post–World War II inflation to reverse itself.

[12] Binder and Brunet (2022) rely on consumer inflation expectations from the Survey of Consumer Finances, a representative survey of consumers that primarily collects information on a household's balance sheet. The survey was conducted annually from 1946 to 1971, in 1977, and every three years since 1983.

Figure 2
Realized and Expected Inflation from 1946 to 2019



Source: Binder and Kamdar (2022).
Note: Consumer Price Index inflation is the year-over-year percent change in the consumer price index for all urban consumers (from FRED series CPIAUCSL, or M04128USM350NNBR before 1948). From the Livingston Survey, we use the median forecast for the Consumer Price Index growth rate from the base period to 12 months ahead (series G_BP_To_12M). From the Survey of Professional Forecasters, we use the 10-year-ahead forecast for Consumer Price Index inflation (series INFCPI10YR). Ten-year inflation breakevens from the Treasury Inflation Protected Securities (TIPS) market since 1999 are from Gürkaynak, Sack, and Wright (2010) (series BKEVEN10). Synthetic TIPS 10-year breakevens before 1999 are from Groen and Middeldorp (2013).

The United States entered the Korean War in June 1950. Later that same year, median inflation expectations reached 4 percent. Inflation and expectations stabilized following the Fed-Treasury Accord of 1951, when an agreement was reached that the Fed would focus on its dual mandate of full employment and low inflation, rather than seeking to accommodate federal borrowing with low interest rates. The newly independent Fed enjoyed strong credibility through the early 1960s (Bordo and Siklos 2014). Inflation expectations of Livingston forecasters from 1952 to 1964 were low and stable, averaging 0.5 percent and ranging from -1.4 to 1.3 percent, while realized inflation averaged 1.3 percent.

The Great Inflation of the 1970s and the Volcker Disinflation

Livingston forecasters' inflation expectations began to rise in the mid-1960s, along with actual inflation, as shown in Figure 2. Bordo and Siklos (2014) argue that the Federal Reserve lost credibility for low inflation in the mid- to late 1960s, when it allowed inflation to creep upwards in order to accommodate the Johnson administration's expansionary fiscal policies—often referred to as the "guns and butter" fiscal policy for pursing both the Vietnam war and expanded social programs.

William McChesney Martin (1969), near the end of his 18-year term as Federal Reserve Chair, reflected that "my term as chairman is ending on a note reminiscent of its beginning. It began with a mighty effort by the Federal Reserve to control the inflation that accompanied the Korean conflict. It is ending with another mighty effort—against the background of another land war in Asia—to control the current inflation and expectations of further inflation." Martin added,

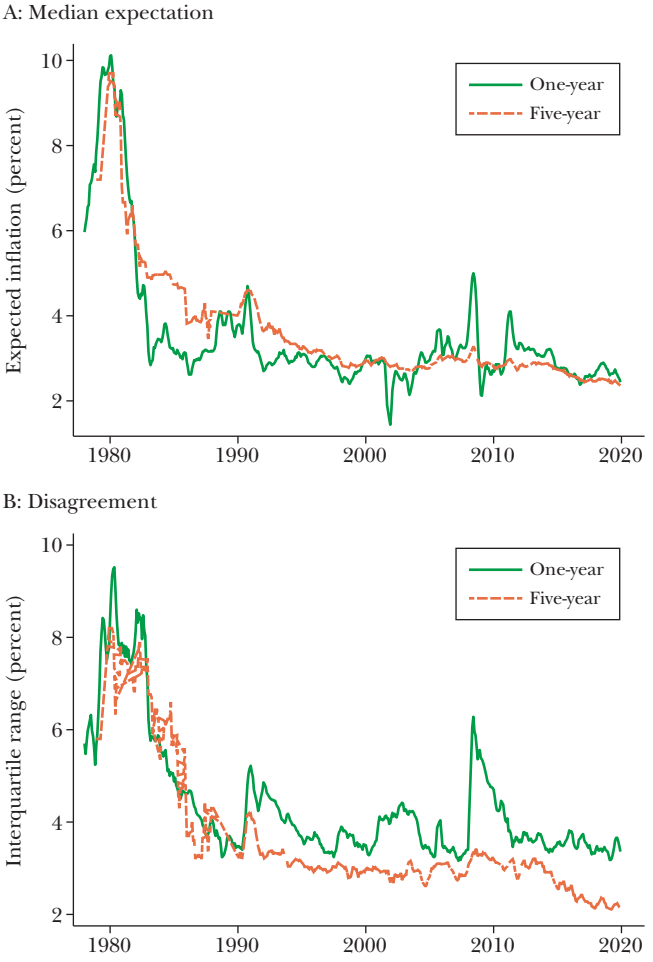
> I believe that we are making progress against the forces that give rise to inflation . . . And we are also, I think, putting some dents in the inflationary expectations that have motivated many of our businesses and consumers. After several years of rapidly rising prices, it is only natural that many spending decisions would be motivated by the fear that prices will be higher next year . . . But there is evidence now, however fragmentary, that these attitudes are changing, however slowly.

In retrospect, Martin was overly optimistic. Under Arthur Burns, Martin's successor as Fed Chair, both actual and expected inflation continued to rise. G. William Miller took over as Fed Chair in 1978, and when Paul Volcker replaced Miller in 1979, he recognized the problematic feedback between rising inflation and rising inflation expectations. Volcker (1979) argued before Congress:

> An entire generation of young adults has grown up since the mid-1960's knowing only inflation, indeed an inflation that has seemed to accelerate inexorably. In the circumstances, it is hardly surprising that many citizens have begun to wonder whether it is realistic to anticipate a return to general price stability, and have begun to change their behavior accordingly. Inflation feeds in part on itself, so part of the job of returning to a more stable and more productive economy must be to break the grip of inflationary expectations.

Goodfriend and King (2005, p. 986) argue that Fed actions in 1979 and 1980 "merely contained inflation in the face of sharply rising inflation expectations." From transcripts of the meetings of the Federal Open Market Committee, they find that the continuation of rising inflation expectations in 1981 finally convinced the Fed to make a more decisive and sustained effort to reduce inflation. Inflation and expected inflation finally began a steady decline, falling below 5 percent in 1982 and below 4 percent in 1983, as the economy exited from a pronounced recession. Figure 2 also displays ten-year-ahead inflation expectations of professional forecasters, as well as inflation expectations based on the ten-year breakevens calculated based on Treasury Inflation Protected Securities (with synthetic values prior to 1999 as discussed in the previous section). The ten-year breakeven measure suggests that inflation expectations fell slowly and were volatile throughout the 1980s and 1990s. However, inflation expectations as derived from ten-year breakeven or from median professional forecasts stabilized by the late 1990s.

Figure 3
Consumer Inflation Expectations and Disagreement, 1978–2019



Source: Binder and Kamdar (2022).
Note: From the Michigan Survey of Consumers, we use the median one-year horizon inflation expectations (series px1_med_all) and median five- to ten-year horizon inflation expectation (series px5_med_all). Disagreement is the interquartile range of inflation expectations from the Michigan Survey of Consumers. For visual clarity, centered five-month moving averages are displayed for each series.

This process of stabilizing inflation expectations is also apparent in the Michigan Survey of Consumers data on inflation expectations. Median consumer inflation expectations at the one-year and five- to ten-year horizon are plotted in panel A of Figure 3. During the 1980s, longer-run inflation expectations fell more slowly than shorter-run expectations, hovering around 5 percent until 1986, and not falling below 4 percent until 1991. Panel B of Figure 3 shows consumer disagreement about

short- and long-run inflation, measured by the cross-sectional interquartile range of inflation expectations. Disagreement rose and then declined with inflation in the late 1970s and early 1980s. Interestingly, longer-run disagreement fell by more than shorter-run disagreement, and remains lower, signaling improved anchoring of long-run household expectations—that is, even when households disagree about inflation over shorter horizons, they are more in agreement about inflation over the longer run.[13]

The Great Inflation of the 1970s and the following Volcker disinflation highlighted the importance of anchoring inflation expectations to stabilize actual inflation, and this lesson has remained highly influential for policymakers to the present day. Federal Reserve Chair Janet Yellen (2015), for instance, has attributed the Great Inflation to the "emergence of an 'inflationary psychology' whereby a rise in actual inflation led people to revise up their expectations for future inflation" which "caused inflation—actual and expected—to ratchet higher over time." Lessons from this episode prompted the widespread adoption of inflation targeting frameworks in the years that followed.

Inflation Targeting before and after the Great Recession

Beginning with the example of the Reserve Bank of New Zealand in 1990, central banks around the world began implementing a monetary policy framework called inflation targeting (for a thorough survey, see Svensson 2010). Inflation targeting involves an announced numerical inflation target and implementation of monetary policy that gives a large role to the inflation forecast, along with transparency and accountability. By 2010, there were roughly 25 inflation-targeting countries. The circumstances for undertaking the policy have varied. For instance, inflation targeting was implemented in New Zealand (with other reforms) following dissatisfaction with the previous government. In the United Kingdom, inflation targeting was adopted in 1992, after being forced away from a fixed exchange rate regime (Pétursson 2005). Despite the varied circumstances under which monetary authorities came to inflation targeting, one goal (either explicit or implicit) of adopting the strategy was to anchor inflation expectations and, in turn, stabilize inflation itself. The early empirical evidence, which relied on case studies or small samples, found that inflation targeting was successful in both goals (Bernanke et al. 1999; Neumann and von Hagen 2002), although subsequent work was less conclusive and pointed to the challenges of identifying the *causal* effects of inflation targeting.[14]

[13] Binder (2017) uses this data to construct an uncertainty index for inflation expectations, based on consumers' propensity to round their forecasts to multiples of five percent. In qualitative terms, this index follows the same general pattern of rising and falling as the disagreement index discussed in the text.

[14] In emerging market economies, inflation targeters experienced lower and less volatile inflation than non-targeters (Gonçalves and Salles 2008; Lin and Ye 2009). For developed countries, there was no significant difference between targeters and non-targeters: both saw a decline in the level and volatility of inflation (Ball and Sheridan 2004). However, as Gertler (2004) notes, classifying advanced economies

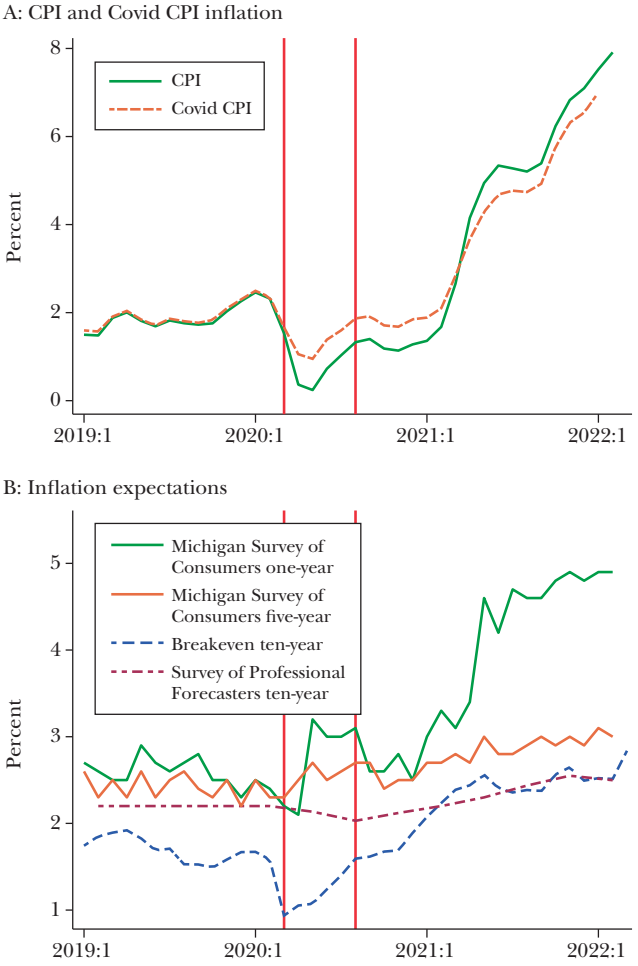
In the United States, discussions about implementing an inflation target began in the mid-1990s, but the Federal Reserve's explicit target of 2 percent inflation as measured by the Personal Consumption Expenditures index was announced only in 2012 (Shapiro and Wilson 2019). As shown earlier in Figure 2, professional forecasters' long-run forecasts for Consumer Price Index inflation had fallen and stabilized near 2.5 percent in 1998. Since Consumer Price Index inflation is around half a percentage point higher than Personal Consumption Expenditures inflation (Binder, Janson, and Verbrugge 2020), due to the different baskets of goods and varying methods of calculating the two indexes, a 2.5 percent Consumer Price Index inflation forecast is consistent with the 2 percent Personal Consumption Expenditures inflation target. Some professional forecasters may have believed that the Fed had an implicit inflation target before the explicit announcement in 2012. In fact, a questionnaire added to the Survey of Professional Forecasters in 2007:IV asked respondents whether they believed the Fed had a numerical target for long-run inflation. Of the 45 respondents, 23 believed that the Fed had such a target.

For consumers, the reaction to inflation targeting appears somewhat different. From Figure 3, see that consumers' longer-run inflation expectations and disagreement trended slightly downward in the years following the 2012 Fed announcement. All remain lower and more stable than the same series for the shorter horizon. However, median longer-run inflation expectations have stayed closer to 3 percent rather than 2 percent, and disagreement among the general population remains much higher than for professional forecasters. Other surveys have explicitly asked consumers whether they know the inflation target, and find that most do not (Binder and Rodrigue 2018; Binder 2020a; Binder 2021). Hence, the 2012 announcement itself may have done relatively little to anchor consumers' expectations. Rather, years of low inflation may have reduced the "inflationary psychology" of earlier decades. As then Fed Chair Yellen (2015) remarked, "Anchored inflation expectations were not won easily or quickly: Experience suggests that it takes many years of carefully conducted monetary policy to alter what households and firms perceive to be inflation's 'normal' behavior."

By the time the Great Recession began in 2007, inflation expectations of consumers and professionals had been low and stable for several years, due in part to implicit inflation targeting and a reduction in inflationary psychology. This anchoring of inflation expectations played a key role in inflation stabilization during and after the Great Recession. Inflation stayed surprisingly stable through the dramatic rise in unemployment in the Great Recession (rather than declining) and again through the recovery (rather than increasing). This weak co-movement between unemployment and inflation became known as the "missing disinflation" and "missing re-inflation" puzzles and prompted proclamations of the death of the original Phillips

into inflation targeting and non-inflation targeting is not a sharp distinction. Many of the non-targeters were either implicitly following an inflation target or had hybrid targets for inflation and money growth. Therefore, the results for advanced economies can be difficult to interpret and may even suggest that inflation targeting has lowered inflation rates and volatility (Svensson 2010).

Figure 4
Inflation and Expected Inflation since 2019



Source: Binder and Kamdar (2022).
Note: Panel A plots Consumer Price Index inflation and Cavallo's (2020) "Covid CPI" inflation. Panel B plots the inflation expectations for several groups: for consumers, a one-year and five-year horizon from the Michigan Survey of Consumers; for financial market participants, from Treasury Inflation Protected Securities ten-year breakevens; and from the Survey of Professional Forecasters, 10-year Consumer Price Index inflation forecasts. There are vertical lines at March 2020, when COVID-19 was declared a pandemic, and August 2020, when "average inflation targeting" was adopted by the Federal Reserve.

curve relationship between inflation and unemployment. However, incorporating anchored long-run inflation expectations can help solve both puzzles and revive the New Keynesian Phillips curve (Jørgensen and Lansing 2019; Hazell et al. 2020). This more recent literature implies that anchored long-term inflation expectations can powerfully stabilize inflation. In our view, this is the most compelling explanation of inflation dynamics from 2008 to 2019.

The Inflation Surge of 2021

At the beginning of the COVID-19 pandemic early in 2020, it was difficult to predict whether the impact would be inflationary or disinflationary, because of the difficulty of distinguishing aggregate supply and aggregate demand shocks (Cochrane 2020). The solid green line in the top panel of Figure 4 shows that inflation as measured by the Consumer Price Index declined in the first few months of the pandemic and then rebounded slightly in the later months of 2020. In mid-2021, as the COVID-19 pandemic continued, aggregate demand recovered but supply remained constrained, and both inflation expectations and inflation itself began to rise.

During this time, how did inflation expectations of different groups adjust? The second panel of Figure 4 shows the evolution of consumer inflation expectations from the Michigan Survey over one- and five-year horizons, professional forecaster inflation expectations from the Survey of Professional Forecasters over the ten-year horizon, and market-implied inflation expectations imputed from Treasury Inflation Protected Securities for a ten-year horizon. Median consumer inflation expectations declined by much less than inflation expectations from the other sources. Why did consumer inflation expectations differ from realized inflation and from the expectations of professional forecasters at the start of the pandemic? At least three explanations have been proposed.

First, the pandemic may have shifted consumption patterns in ways that led to consumers experiencing price pressures that differed from the basket of goods behind the Consumer Price Index (CPI). Cavallo (2020) calculated a "Covid CPI" series, which uses credit and debit card transaction data to adjust the weights in the basket of goods underlying the Consumer Price Index in order to match the new expenditure patterns. For the first year of the pandemic, Covid CPI inflation was higher than official Consumer Price Index inflation. Thus, consumers were experiencing inflation at a higher rate than represented by the official Consumer Price Index, and consumers have been shown to rely heavily on their experienced price changes when forming their expectations (Cavallo, Cruces, and Perez-Truglia 2017). In contrast, professional forecasters and financial market participants tend to rely on official information sources.

A second possible explanation for the discrepancy between consumer and professional forecaster expectations, especially during the early disinflationary part of the pandemic, is that consumers may not distinguish between aggregate supply and aggregate demand shocks, but instead may simply associate bad news with high inflation (Binder and Makridis 2022; Binder 2020b; Kamdar 2019). Thus, consumers' expectations can greatly diverge from actual inflation and from professional forecasters' expectations when there is an adverse aggregate demand shock, as in the Great Recession and early stages of the pandemic. Binder (2020a) surveyed consumers on March 5 and 6, 2020, shortly after the Fed's emergency rate cut on March 3 (which 38 percent of survey respondents knew about). Consumers who were more concerned about the pandemic had significantly higher inflation

expectations, consistent with prior research showing that many consumers seem to lack a clear understanding of the drivers of inflation or of the role of monetary policy.

Third, the inflation expectations of consumers may be less responsive to the release of official macroeconomic news than the expectations of professional forecasters and market participants. For market participants, it is possible to test how inflation expectations respond to macroeconomic news and announcements using an event-study approach, since market data on Treasury Inflation-Protected Securities is available at daily frequency. Bauer (2015) shows that inflation compensation in this data responds to macroeconomic data surprises, including surprises to the "core" Consumer Price Index (which leaves out prices changes in the volatile food and energy categories). For consumers, daily data on inflation expectations is generally not available, though some researchers have conducted their own daily surveys around announcements of interest. Monetary policy announcements seem to have minimal effects on consumer inflation expectations (Lamla and Vinogradov 2019), and the June 2021 release of the Consumer Price Index, which came in surprisingly high, only affected the inflation expectations of highly numerate consumers (Binder 2021).

In addition to the COVID-19 pandemic, another development that may have affected inflation expectations and inflation in recent years was the August 2020 announcement by the Federal Reserve that it would adjust its policy framework from an "inflation targeting" approach previously announced in January 2012 to an "average inflation targeting" approach. The shift in terminology implies that if inflation is below the target level for some time, then it will be allowed to rise above target in the future, and vice versa. In discussions of average inflation targeting, Fed officials emphasize "the importance of having well-anchored inflation expectations, both to foster price stability and to enhance our ability to promote our broad-based and inclusive maximum-employment goal" (Powell 2021).

Like the January 2012 inflation targeting announcement, the average inflation targeting announcement did not have drastic immediate effects on expectations. Rather, inflation expectations, especially at longer horizons, rose gradually with inflation itself. Data from the market for Treasury Inflation Protected Securities implied ten-year inflation expectations steadily rose from 1 percent in May 2020 to 2.4 percent in July 2021, while the median professional forecast from the Survey of Professional Forecasters rose from 1.9 percent to 2.4 percent. The Michigan Survey of Consumers' expectations also rose with realized inflation beginning in late 2020, especially at the one-year horizon. These short-run expectations are at 4.9 percent as of February 2022. Consumer inflation disagreement also rose and remains elevated.

Given the recent increases in expectations and realized inflation, monetary policymakers must consider the risk of inflation expectations becoming unanchored if inflation remains elevated for an extended period. For professional forecasters, the microdata has shown some evidence of weakening anchoring. Binder, Janson, and Verbrugge (2021) suggest a measure of expectations anchoring based on the deviations of individual forecasters' long-run inflation expectations from target over a rolling window. This measure declined—implying improved anchoring—from 2012

until 2018. But in time windows that include the pandemic or rising inflation, forecasters are increasingly reporting long-run inflation forecasts that are further from target, even as the median forecast remains relatively close to target. For consumers, short-run inflation expectations have sharply risen. Consumers have been shown to be more attentive to inflation when inflation is high (Coibion et al. 2020). This raises the prospect that consumers may become more attentive and their long-run inflation expectations could become less anchored in a high inflation environment. Overall, if long-run inflation expectations of firms and consumers increase, the possibility of persistently higher inflation will rise.

As policymakers move to reduce inflation, higher and unanchored inflation expectations could complicate their task. Accordingly, policymakers should pay careful attention to developments in inflation expectations. We suggest a particular focus on the level of long-run inflation expectations, as well as on the range of disagreement for such expectations.

References

Adam, Klaus, and Mario Padula. 2011. "Inflation Dynamics and Subjective Expectations in the United States." *Economic Inquiry* 49 (1): 13–25.

Allen, William R. 1977. "Irving Fisher, F.D.R., and the Great Depression." *History of Political Economy* 9 (4): 560–587.

Ang, Andrew, Geert Bekaert, and Min Wei. 2007. "Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?" *Journal of Monetary Economics* 54 (4): 1163–212.

Armantier, Olivier, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar. 2016. "An Overview of the Survey of Consumer Expectations." Staff Reports 800, Federal Reserve Bank of New York.

Ball, Laurence, and Niamh Sheridan. 2004. "Does Inflation Targeting Matter?" In *The Inflation-Targeting Debate*, edited by Ben S. Bernanke and Michael Woodford, 249–82. Chicago: University of Chicago Press.

Bauer, Michael. 2015. "Inflation Expectations and the News." *International Journal of Central Banking* 11 (2): 1–40.

Bernanke, Ben S. 1983. "Irreversibility, Uncertainty, and Cyclical Investment." *Quarterly Journal of Economics* 98 (1): 85–106.

Bernanke, Ben S., Thomas Laubach, Frederic S. Mishkin, and Adam S. Posen. 1999. *Inflation Targeting: Lessons from the International Experience*. Princeton: Princeton University Press.

Binder, Carola. 2016. "Estimation of Historical Inflation Expectations." *Explorations in Economic History* 61 (C):1–31.

Binder, Carola. 2017. "Measuring Uncertainty Based on Rounding: New Method and Application to Inflation Expectations." *Journal of Monetary Economics* 90 (C): 1–12.

Binder, Carola. 2020a. "Coronavirus Fears and Macroeconomic Expectations." *Review of Economics and Statistics* 102 (4): 721–30.

Binder, Carola. 2020b. "Long-run Inflation Expectations in the Shrinking Upper Tail." *Economics Letters* 186 (C).

Binder, Carola. 2021. "Household Expectations and the Release of Macroeconomic Statistics." *Economics Letters* 207 (110041).

Binder, Carola, and Gillian Brunet. 2022. "Inflation Expectations and Consumption: Evidence from 1951." *Economic Inquiry* 60 (2): 954–74.

Binder, Carola, Wesley Janson, and Randal Verbrugge. 2020. "The CPI–PCEPI Inflation Differential: Causes and Prospects." *Economic Commentary*, March 2. https://www.clevelandfed.org/newsroom-and-events/publications/economic-commentary/2020-economic-commentaries/ec-202006-cpi-pcepi-inflation-differential.aspx.

Binder, Carola, Wesley Janson, and Randal Verbrugge. 2021. "Out of Bounds: Do SPF Respondents Have Anchored Inflation Expectations?" Unpublished.

Binder, Carola, and Rupal Kamdar. 2022. "Replication data for: Expected and Realized Inflation in Historical Perspective." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E166881V1.

Binder, Carola, and Christos Makridis. 2022. "Stuck in the Seventies: Gas Prices and Consumer Sentiment." *Review of Economics and Statistics* 104 (2): 293–305.

Binder, Carola, and Alex Rodrigue. 2018. "Household Informedness and Long-Run Inflation Expectations: Experimental Evidence." *Southern Economic Journal* 85 (2): 580–98.

Board of Governors of the Federal Reserve System. 1943. *Thirtieth Annual Report of the Board of Governors of the Federal Reserve System: Covering Operations for the Year 1943*. Washington, DC: Board of Governors of the Federal Reserve System.

Bordo, Michael D. 1981. "The Classical Gold Standard: Some Lessons for Today." *Federal Reserve Bank of St. Louis* 63: 1–17.

Bordo, Michael, and Pierre Siklos. 2014. "Central Bank Credibility, Reputation and Inflation Targeting in Historical Perspective." NBER Working Paper 20693.

Cagan, Phillip. 1956. "The Monetary Dynamics of Hyperinflation." In *Studies in the Quantity Theory of Money*, edited by Milton Friedman. Chicago: University of Chicago Press.

Cagan, Phillip. 1965. *Determinants and Effects of Changes in the Stock of Money, 1875–1960*. Cambridge, MA: National Bureau of Economic Research.

Calvo, Guillermo A. 1983. "Staggered Prices in a Utility-Maximizing Framework." *Journal of Monetary Economics* 12 (3): 383–98.

Cavallo, Alberto. 2020. "Inflation with COVID Consumption Baskets." Harvard Business School Working Paper 20-124.

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia. 2017. "Inflation Expectations, Learning, and Supermarket Prices: Evidence from Survey Experiments." *American Economic Journal: Macroeconomics* 9 (3): 1–35.

Cecchetti, Stephen G. 1992. "Prices During the Great Depression: Was the Deflation of 1930–1932 Really Unanticipated?" *American Economic Review* 82 (1): 141–56.

Clark, Todd E., and Troy A. Davig. 2008. "An Empirical Assessment of the Relationships among Inflation and Short- and Long-term Expectations." Federal Reserve Bank of Kansas City Research Working Paper 08-05.

Clark, Todd E., and Troy Davig. 2009. "Memo: The Relationship between Inflation and Inflation Expectations," in United States. Federal Open Market Committee, Federal Open Market Committee Meeting Minutes, Transcripts, and Other Documents (December 15–16, 2009): 1–17.

Clark, Todd E., and Taisuke Nakata. 2008. "Has the Behavior of Inflation and Long-Term Inflation Expectations Changed?" *Economic Review* [Federal Reserve Bank of Kansas City] 93 (1): 17–50.

Cochrane, John H. 2020. "Coronavirus Monetary Policy." In *Economics in the Time of COVID-19*, edited by Richard Baldwin and Beatrice Weder di Mauro, 105–8. London: CEPR Press. https://voxeu.org/content/economics-time-covid-19.

Cogley, Timothy, and Argia M. Sbordone. 2008. "Trend Inflation, Indexation, and Inflation Persistence Coibion, Olivier, and Yuriy Gorodnichenko. 2012. "What Can Survey Forecasts Tell Us about Information Rigidities?" *Journal of Political Economy* 120 (1): 116–59.

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

Coibion, Olivier, Yuriy Gorodnichenko, and Rupal Kamdar. 2018. "The Formation of Expectations, Inflation, and the Phillips Curve." *Journal of Economic Literature* 56 (4): 1447–91.

Coibion, Olivier, Yuriy Gorodnichenko, and Saten Kumar. 2018. "How Do Firms Form Their Expectations? New Survey Evidence." *American Economic Review* 108 (9): 2671–713.

Coibion, Olivier, Yuriy Gorodnichenko, Saten Kumar, and Mathieu Pedemonte. 2020. "Inflation Expectations as a Policy Tool?" *Journal of International Economics* 124: 103297.

Croushore, Dean D. 1997. "The Livingston Survey: Still Useful After All These Years." *Business*

Review-Federal Reserve Bank of Philadelphia 2: 1–12.
Curtin, Richard. 2016. "George Katona: A Founder of Behavioral Economics." In *Routledge Handbook of Behavioral Economics*, edited by Roger Frantz, Shu-Heng Chen, Kurt Dopfer, Floris Heukelom, and Shabnam Mousavi. Routledge, 30–47. Abingdon: Routledge.
Dimand, Robert W. 1999. "Irving Fisher and the Fisher Relation: Setting the Record Straight." *Canadian Journal of Economics* 32: (3): 744–50.
Dorval, Bill, and Gregor W. Smith. 2013. "Interwar Deflation and Depression." Queen's Economics Department Working Paper 1310.
Eggertsson, Gauti B. 2008. "Great Expectations and the End of the Depression." *American Economic Review* 98 (4): 1476–516.
Fisher, Irving. 1911. *The Purchasing Power of Money: Its Determination and Relation to Credit, Interest, and Crises*. New York: Macmillan Company.
Fisher, Irving. 1930. *The Theory of Interest*. New York: Macmillan Company.
Fisher, Irving. 1933. "The Debt Deflation Theory of Great Depressions." *Econometrica* 1 (4): 337–57.
Friedman, Milton. 1968. "The Role of Monetary Policy." *American Economic Review* 58: 1–17.
Friedman, Milton, and Anna Jacobson Schwartz. 1963. *A Monetary History of the United States, 1867–71960*. Princeton: Princeton University Press.
Galí, Jordi. 2018. "The State of New Keynesian Economics: A Partial Assessment." *Journal of Economic Perspectives* 32 (3): 87–112.
Gertler, Mark. 2004. "Comment: Does Inflation Targeting Matter?" In *The Inflation-Targeting Debate*, edited by Ben S. Bernanke and Michael Woodford, 276–281. Chicago: University of Chicago Press.
Ginsburg, David. 1952. Price Stabilization, 1950–1952: Retrospect and Prospect. *University of Pennsylvania Law Review* 100 (4): 514–43.
Goodfriend, Marvin, and Robert G. King. 2005. "The Incredible Volcker Disinflation." *Journal of Monetary Economics* 52: 981–1015.
Gonçalves, Carlos Eduardo S., and João M. Salles. 2008. "Inflation Targeting in Emerging Economies: What do the Data Say?" *Journal of Development Economics* 85 (1–2): 312–18.
Gordon, Robert J. 2018. "Friedman and Phelps on the Phillips Curve Viewed from a Half Century's Perspective." *Review of Keynesian Economics* 6 (4): 425–36.
Groen, Jan J. J., and Menno Middeldorp. 2013. "Creating a History of U.S. Inflation Expectations." *Liberty Street Economics*, August 21. https://libertystreeteconomics.newyorkfed.org/2013/08/creating-a-history-of-us-inflation-expectations/.
Gürkaynak, Refet S., Brian Sack, and Jonathan H. Wright. 2010. "The TIPS Yield Curve and Inflation Compensation." *American Economic Journal: Macroeconomics* 2 (1): 70–92.
Hall, Robert E., and Thomas J. Sargent. 2018. "Short-Run and Long-Run Effects of Milton Friedman's Presidential Address." *Journal of Economic Perspectives* 32 (1): 121–34.
Hamilton, James D. 1992. "Was the Deflation During the Great Depression Anticipated? Evidence from the Commodity Futures Market." *American Economic Review* 82 (1): 157–78.
Hazell, Jonathon, Juan Herreño, Emi Nakamura, and Jón Steinsson. 2020. "The Slope of the Phillips Curve: Evidence from US States." NBER Working Paper 28005.
Jalil, Andrew J., and Gisela Rua. 2016. "Inflation Expectations and Recovery in Spring 1933." *Explorations in Economic History* 62: 26–50.
Jørgensen, Peter, and Kevin J. Lansing. 2019. "Anchored Inflation Expectations and the Flatter Phillips Curve." Federal Reserve Bank of San Francisco Working Paper Series 2019-27.
Kamdar, Rupal. 2019. "Inattentive Consumer: Sentiment and Expectations." Unpublished
Klein, Benjamin. 1975. "Our New Monetary Standard: The Measurement and Effects of Price Uncertainty, 1880–1973." *Economic Inquiry* pages 461–84.
Lamla, Michael J., and Dmitri V. Vinogradov. 2019. "Central Bank Announcements: Big News for Little People?" *Journal of Monetary Economics* 108: 21–38.
Lin, Shu, and Haichun Ye. 2009. "Does Inflation Targeting Make a Difference in Developing Countries?" *Journal of Development Economics* 89 (1): 118–23.
Lowe, Joseph. 1823. *On the Present State of England*. London: Longman, Hurst, Rees, Orme, and Brown.
Lucas, Robert E. 1972. "Expectations and the Neutrality of Money." *Journal of Economic Theory* 4 (2): 103–24.
Lucas, Robert E. 1973. "Some International Evidence on Output-Inflation Tradeoffs." *American Economic Review* 63 (3): 326–34.
Martin, Jr., William McChesney. 1969. "Reminiscences and Reflections: Remarks before The Business

Council, Hot Springs, Virginia." Board of Governors of the Federal Reserve System, October 17. https://fraser.stlouisfed.org/title/statements-speeches-william-mcchesney-martin-jr-448/reminiscences-reflections-7946/.

Muth, John F. 1961. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (3): 315–35.

Nelson, Daniel B. 1991. "Was the Deflation of 1929-1930 Anticipated? The Monetary Regime as Viewed by the Business Press." *Research in Economic History* 13: 1–65.

Neumann, Manfred J.M., and Jürgen von Hagen. 2002. "Does Inflation Targeting Matter?" *Federal Reserve Bank of St. Louis Review* 84 (4): 127–48.

Persky, Joseph. 1998. "Price Indexes and General Exchange Values." *Journal of Economic Perspectives* 12 (1): 197–205.

Pétursson, Thórarinn G. 2005. *Inflation Targeting and Its Effects on Macroeconomic Performance*. SUERF Studies 2005/5. Vienna: SUERF.

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

Powell, Jerome. 2019. "Oversight on The Monetary Policy Report to Congress Pursuant to The Full Employment and Balanced Growth Act of 1978." US Senate Committee on Banking, Housing, and Urban Affairs, February 26. https://www.govinfo.gov/content/pkg/CHRG-116shrg35838/html/CHRG-116shrg35838.htm.

Powell, Jerome. 2021. "Transcript of Chair Powell's Press Conference." Federal Reserve Board of Governors, September 22, https://www.federalreserve.gov/mediacenter/files/FOMCpresconf20210922.pdf.

Rockoff, Hugh. 1981. "Price and Wage Controls in Four Wartime Periods." *Journal of Economic History* 41 (2): 381–401.

Romer, Christina D., and David H. Romer. 2013. "The Missing Transmission Mechanism in the Monetary Explanation of the Great Depression." *American Economic Review* 103 (3): 66–72.

Rudd, Jeremy B. 2021. "Why Do We Think That Inflation Expectations Matter for Inflation? (And Should We?)." Finance and Economics Discussion Series 2021-062. Washington: Board of Governors of the Federal Reserve System. https://doi.org/10.17016/FEDS.2021.062.

Sargent, Thomas J. 1973. "Interest Rates and Prices in the Long Run: A Study of the Gibson Paradox." *Journal of Money, Credit and Banking* 5 (1): 385–449.

Sargent, Thomas J. 1980. "Rational Expectations and the Reconstruction of Macroeconomics." Federal Reserve Bank of Minneapolis Quarterly Review 4 (3): 15–19.

Sargent, Thomas J., and Neil Wallace. 1976. "Rational Expectations and the Theory of Economic Policy." *Journal of Monetary Economics* 2 (2): 169–83.

Schwartz, Anna J. 1981. "Understanding 1929–1933." In *The Great Depression Revisited*, edited by Karl Brunner, 5–48. Boston: Martinus Nijhoff.

Shapiro, Adam, and Daniel J. Wilson. 2019. "The Evolution of the FOMC's Explicit Inflation Target." FRBSF Economic Letter 12, Federal Reserve Bank of San Francisco.

Svensson, Lars E. O. 2010. "Inflation Targeting." In *Handbook of Monetary Economics*, Vol. 3, edited by Benjamin M. Friedman and Michael Woodford, 1237–302. San Diego: Elsevier.

Temin, Peter, and Barrie A. Wigmore. 1990. "The End of One Big Deflation." *Explorations in Economic History*, 27 (4): 483–502.

Thomas, Lloyd B. 1999. "Survey Measures of Expected US Inflation." *Journal of Economic Perspectives* 13 (4): 125–44.

Volcker, Paul. 1979. "Statement before the Joint Economic Committee of the US Congress." Board of Governors of the Federal Reserve System, October 17. https://fraser.stlouisfed.org/title/statements-speeches-paul-a-volcker-451/statement-joint-economic-committee-8205.

Yellen, Janet. 2015. "Inflation Dynamics and Monetary Policy." Speech, Philip Gamble Memorial Lecture, University of Massachusetts, Amherst, September 24. https://www.federalreserve.gov/newsevents/speech/yellen20150924a.htm.

Zarnowitz, Victor, and Phillip Braun. 1993. "Twenty-Two Years of the NBER-ASA Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance." In *Business Cycles, Indicators, and Forecasting*, edited by James H. Stock and Mark W. Watson, 11–94. Chicago: University of Chicago Press.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 157–184

The Subjective Inflation Expectations of Households and Firms: Measurement, Determinants, and Implications

Michael Weber, Francesco D'Acunto, Yuriy Gorodnichenko, and Olivier Coibion

Federal Reserve chair Jerome Powell (2021) recently said: "Inflation expectations are terribly important. We spend a lot of time watching them." Why would these expectations be so important? The traditional policy view is that inflation expectations help central banks and other institutions predict future inflation rates and hence feed into the production of economic forecasts—one of the main tasks policy institutions perform. Indeed, the survey-based inflation expectations of professionals and households have been shown to help forecast future inflation (Ang, Bekaert, and Wei 2007). Traditionally, macroeconomic researchers have also stressed an important role for the inflation expectations of a specific group of agents, financial market participants, because such expectations have been shown to affect asset prices, such as stock prices and interest rates (Bernanke and Kuttner 2005).

■ *Michael Weber is Associate Professor of Finance, Booth School of Business, University of Chicago, Chicago, Illinois. Francesco D'Acunto is James A. Clark Associate Professor of Finance, Economics, and Public Policy, McDonough School of Business, Georgetown University. Yuriy Gorodnichenko is Quantedge Presidential Professor of Economics, University of California-Berkeley, Berkeley, California. Olivier Coibion is Malcolm Forsman Centennial Professor of Economics, University of Texas, Austin, Texas. Weber is a Faculty Research Fellow and Gorodnichenko and Coibion are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Ordering of authors has been randomly selected. Their email addresses are michael.weber@chicagobooth.edu, dacuntof@bc.edu, ygorodni@econ.berkeley.edu, and ocoibion@econ.utexas.edu.*

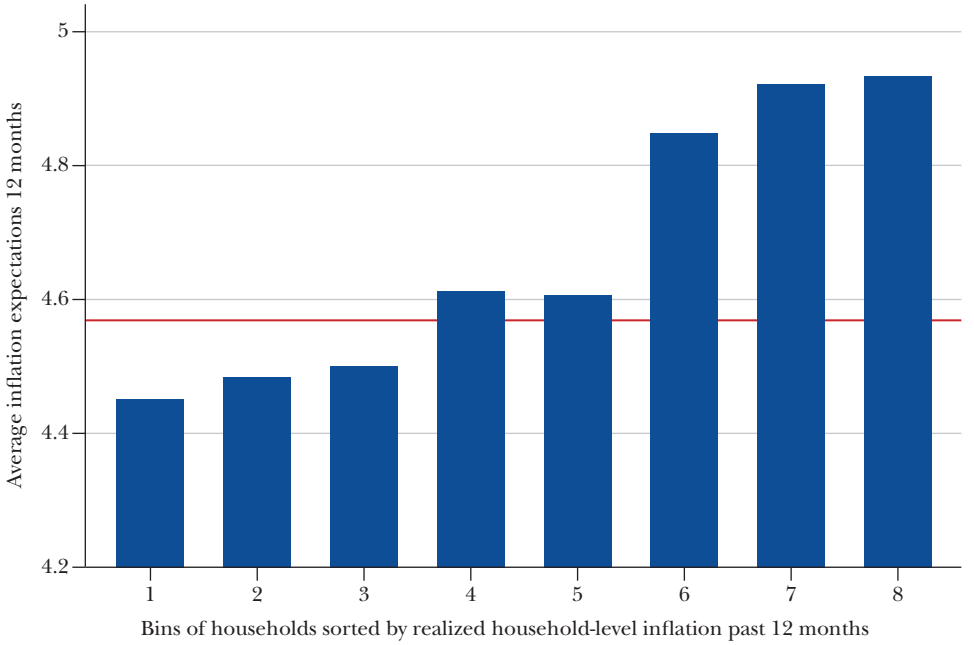
For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.157.

These two traditional roles of inflation expectations, though, are *not* the ones central bankers such as Powell have been emphasizing since the Great Recession. In their view, the key reason why subjective inflation expectations matter is that they affect the prices and wages firms set as well as the consumption-saving decisions of households. This view does not focus on the expectations of financial-market participants or professional forecasters—of which most firms and households are barely ever aware—but on the subjective inflation expectations of ordinary economic agents. James Bullard (2016), President of the St. Louis Federal Reserve Bank, laid out this logic clearly in explaining why inflation expectations are important: "Firms and households take into account the expected rate of inflation when making economic decisions, such as wage contract negotiations or firms' pricing decisions." If subjective inflation expectations affect such important choices for individual and aggregate outcomes, understanding the patterns of inflation expectations in the cross-section and time series is crucial. The driving forces behind their heterogeneity across individuals and firms can also help us understand why otherwise similar economic decision-makers react so differently to the same business-cycle shocks and policy interventions, patterns that traditional representative-agent models cannot capture.

Why would households and firms take their subjective inflation expectations into account when making fundamental economic choices? In theory, how rapidly households expect prices to increase in the future should matter for how they allocate their spending over time. For example, expectations of much higher prices in the future should induce households to purchase more goods today while prices are still relatively low ("intertemporal substitution"). Also, because nominal prices and wages change only infrequently, high rates of inflation erode the value of sticky nominal prices and wages over time, a feature firms and workers take into account when setting prices as well as when bargaining over wage increases. Subjective inflation expectations also shape expectations of how expensive it will be to repay loans with future dollars, and such expectations are crucial to firms' investment decisions—which typically require external financing—as well as households' choices about how to finance the purchase of large-ticket items such as houses, cars, and other durable goods.

Despite this prominent role of subjective inflation expectations in theoretical models and the assessments of policymakers, economists still know little about how such expectations are formed and why they are so heterogeneous even across agents who appear similar based on demographic characteristics. In fact, even the ways in which subjective inflation expectations (and macroeconomic expectations in general) can be best elicited from a population of agents who are often not economically, financially, or mathematically sophisticated is still an open debate in the profession. A few facts, though, hold systematically across space and over time. For instance, on average the inflation expectations of households and firms are higher than what inflation turns out to be and the disagreement across households and firms is orders of magnitudes higher than that among professional forecasters. Understanding the causes and consequences of these distortions in the beliefs of ordinary agents relative to the inflation that later occurs has been at the center of a recently burgeoning academic literature at the intersection of economics, psychology, marketing, and related fields.

Figure 1
Personal Grocery Inflation and Inflation Expectations



Source: Weber et al. (2022).
Note: This figure plots average individual survey inflation expectations from the Chicago Booth Expectations and Attitudes Survey on the *y*-axis for eight groups of respondents sorted based on the inflation of their personal grocery bundle in the 12 months before the survey (*x*-axis). The *x*-axis contains bins of households with each bin containing approximately 6,250 respondents.

One of the lessons from this literature is that ordinary agents consider the specific economic signals they observe in their own environment, such as the prices households see while shopping or the prices firms see their competitors set, to form and update their own inflation expectations. Figure 1 illustrates this point. This graph plots average individual-level one-year ahead inflation expectations (*y*-axis) from the Chicago Booth Expectations and Attitudes Survey (see Nielsen IQ 2017) against bins of household-specific grocery inflation over the previous 12-month (*x*-axis) (Kaplan and Schulhofer-Wohl 2017; D'Acunto, Malmendier, Ospina, et al. 2021). The inflation agents have observed in their own grocery bundles is indeed correlated with their expectations about future inflation. The differences in average inflation expectations across the extreme bins is large—it amounts to about 0.5 percentage points for a period in which realized inflation was systematically below the 2 percent inflation target by the Federal Reserve.

The presence of systematic associations in the data is prima facie evidence that elicited subjective expectations are not pure noise: if they were, we would not detect systematic patterns. Yet eliciting and measuring the inflation expectations of agents

who typically know relatively little about economics poses daunting challenges and stumbling blocks. Survey respondents will always provide an answer when forced, but whether such an answer truly reflects actual beliefs will depend on whether agents understand survey questions, on the ways in which agents conceptualize inflation and other macroeconomic variables, and on the effort agents put in forming beliefs when asked, given that (contrary to the case of professional forecasters) providing accurate inflation expectations can barely be incentivized in a survey of ordinary agents. Reassuringly, researchers have learned a lot about how to design surveys for firms and households that can provide high-quality measures of their expectations about subsequent price changes—so much so that surveys of expected inflation often span decades and are available in dozens of countries.

Once consistent facts are established across space and over time, the question of how we should interpret such facts becomes compelling. Interpreting facts is ultimately a quest for the deep-rooted and underlying determinants of subjective beliefs. For instance, at least since Lucas (1972), economists have conjectured that the signals about price changes agents see explicitly around them should shape their subjective inflation expectations. Following this line of reasoning, grocery and gasoline price changes should play a particularly important role, because ordinary households observe such prices frequently in their daily lives. Relying on personal signals about price changes might also help explain the observed heterogeneity and dispersion of subjective inflation expectations, because agents purchase different bundles of goods, shop at different outlets, and engage with different sets of suppliers and customers, and hence they observe different price changes, which can feed into conflicting views about the likely path of future prices.

Ultimately, we care about subjective inflation expectations only to the extent that such expectations can help us understand heterogeneous choices and reactions observed in the data after the same shocks and policy interventions. Recent macroeconomic research using individual-level transaction data has demonstrated that subjective inflation expectations do explain heterogeneous economic decisions at the individual level and also shape macroeconomic aggregates.

The wealth of new and recent data on inflation expectations and individual-level economic choices of households and firms makes these research endeavors increasingly feasible and compelling and an exciting frontier for researchers in empirical macroeconomics, behavioral economics, finance, marketing, cognitive sciences, and many related fields. In our online Appendix, we list more than 49 survey-based sources of individual-level subjective inflation expectations that have become available across a number of countries over the last few years, with weblinks for each. These represent a wealth of data for researchers interested in the study of subjective inflation expectations.[1]

[1] Online Appendix Table 1 contains details on how to access microdata on inflation expectations for firms in Italy and the United States and on inflation expectations for households in the following countries: Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, Colombia, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Germany, India, Indonesia, Ireland, Italy, Greece, Hungary, France, Japan,

Challenges for Measuring the Inflation Expectations of Households and Firms

One might think that eliciting subjective inflation expectations through surveys is simple: just ask a representative sample what they think inflation will be over some horizon and then record this truthful, informed, and unbiased response. In reality, researchers have to wrestle with a number of challenges. Some of these challenges are common to the elicitation of expectations of any kind and some are specific to the measurement of inflation expectations. We highlight the issues we find most concerning in terms of survey design through the lens of the most commonly used US surveys in the literature: the Michigan Survey of Consumers and the New York Fed Survey of Consumer Expectations. For the latter survey, we refer the reader to Armantier et al. (2013) for a comprehensive description and discussion.

Question Wording

The wording of the survey question aimed at eliciting inflation expectations already poses challenges. For instance, the Michigan Survey of Consumers asks households to report their point prediction for the *change in the general level of prices*: "During the next 12 months, do you think that prices in general will go up, or go down, or stay where they are now?" In contrast, the Survey of Consumer Expectations, run by the Federal Reserve Bank of New York, asks households to report their expectations for *inflation*: "Now we would like you to think about the different things that may happen to inflation over the next 12 months." While inflation and the change in the general level of prices may seem equivalent to economists, when asked, the general population, which typically lacks economic and financial literacy, might think about the prices in their nondurable consumption bundle rather than about the overall representative consumption bundle, might confuse levels with changes, or might be unfamiliar with the concept of inflation and have trouble using percentages (for example, see Bruine de Bruin et al. 2012).

Furthermore, neither of the surveys specifies which price index respondents should have in mind when reporting their expectations. This ambiguity allows researchers to reach a higher response rate, in part because respondents are less likely to answer "I don't know" because they are unaware of a specific price index, but it might introduce more disagreement in survey responses. For example, respondents might form their expectations while also considering stock-market prices, which are not a part of the Consumer Price Index or other conventional price indices (Kumar et al. 2015). Some surveys do ask respondents to report their predictions for a specific price index (for example, Coibion et al. 2020), but this approach implicitly assumes that the respondents know the index and its definition.

Latvia, Lithuania, Luxembourg, Malaysia, Malta, Mexico, Montenegro, Netherlands, New Zealand, Philippines, Poland, Portugal, Republic of North Macedonia, Serbia, Singapore, Slovakia, Slovenia, Sweden, Switzerland, Spain, Turkey, the United Kingdom, and the United States.

The use of screener questions that exclude respondents who are illiterate about inflation from the survey pool have been proposed as a potential solution to this challenge, even though they open an issue of selection. The Reserve Bank of New Zealand, for instance, while collecting households' inflation expectations, uses a screener question ("What is your understanding of the term inflation?") to exclude respondents who do not understand the concept of inflation.[2]

Priming

Survey design can inadvertently nudge or "prime" respondents to tilt their answers in a particular direction. For example, if a respondent reports an inflation forecast that an interviewer finds unrealistic, the interviewer may probe the respondent with a clarifying question, which in turn may lead the respondent to adjust toward a "more realistic" value. For example, the Michigan Survey of Consumers provides this instruction to interviewers, "IF R GIVES AN ANSWER THAT IS GREATER THAN 5%, PLEASE PROBE WITH: 'Let me make sure I have that correct. You said that you expect prices to go (up/down) during the next 12 months by (X) percent. Is that correct?'" If probing only happens when respondents provide seemingly unrealistic forecasts of inflation, the elicitation procedure faces an undesirable asymmetry. Follow-up probing questions are meant to reduce noise in survey responses, but they may also lead to a distorted measure of what people truly think about future inflation.

Priming can take a variety of forms. For example, some surveys provide background information like levels of recent inflation. The Survey of Inflation and Growth Expectations, run by the Bank of Italy, tells managers the most recent inflation rate before asking them to report their inflation expectations: "The last [month] consumer price inflation, measured by the 12-month change in the harmonized index of consumer prices was equal to [IT] in Italy and to [EA] in the euro area. What do you think it will be in Italy?" The provision of background information affects the level and dispersion of inflation forecasts (Coibion, Gorodnichenko, and Ropele 2020).

Pre-set answer options and limited lists of possible options can also prime respondents. For instance, the Business Outlook Survey, run by the Bank of Canada, offers only four possible outcomes for inflation forecasts: "less than 1%," "1 to 2%," "2% to 3%," and "more than 3%." Coibion et al. (2020b) document that offering a limited set of choices reduces the dispersion of reported responses. Furthermore, a respondent who is uncertain about future inflation may just pick the center of the provided range if the answer is not open-ended.

Sampling

We live in an age of declining survey response rates (Bureau of Labor Statistics 2021), in part because communication has evolved in ways that bring people to pay less attention to phone calls and physical mail. In this context, reaching out to a representative group of the population and acquiring their consent to participate in

[2]For more details see https://www.rbnz.govt.nz/statistics/m13.

a survey is not easy. Online/computer-based surveys offer the greatest flexibility and can be straightforward for the computer-literate, young, and educated respondent, but often pose barriers for older individuals and those who may be less versed in technology or who evince greater mistrust from automated algorithms (D'Acunto and Rossi 2021). As a result, surveys often have to rely on a mixture of modes—online, phone, in-person—to be representative.

The opportunity cost of participating in a survey also affects enrollment. This issue is particularly stark for surveys of business executives whose time is scarce and who are only indirectly accessible through assistants. As a result, firm surveys of inflation expectations and other topics are often based on convenience samples developed via client lists, club/association members, personal contacts, and so on. In short, finding a typical and representative set of survey respondents can be difficult.

Panel Conditioning

Surveys often attempt to get participants to enroll across multiple waves. Repeated participation can be useful: for example, by looking at the evolution in views of a common set of individuals, selection due to a changing composition is not an issue. But a potential limitation of repeatedly surveying the same individuals about the same topic is that respondents may learn, from their very participation in the survey, about the topic. This effect is commonly known as "panel conditioning." The effect is typically small in most contexts, but in the case of inflation expectations of households and firms, recent evidence indicates that it can be quite large. Kim and Binder (2021) document that households participating in the Survey of Consumer Expectations from the Federal Reserve Bank of New York reduce their inflation expectations by 2 percentage points on average after participating for a few months, which suggests that repeat participants may no longer be considered representative of the broader population.

Point Predictions versus Distributions

Manski (2004) popularized the use of survey questions that elicit subjective probability distributions about future outcomes at the micro and macro levels. For example, the Survey of Consumer Expectations from the Federal Reserve Bank of New York asks respondents to assign probabilities to ten possible ranges of future inflation: for example, "the rate of inflation will be 12% or higher," "the rate of inflation will be between 8% and 12%," and ranging to "the rate of deflation (opposite of inflation) will be 12% or higher." One can use the reported probability distributions to infer not only a central tendency (like mean or mode), which is highly correlated with point forecasts, but also to capture the associated uncertainty in an individual's forecast which can signal precautionary behavior resulting in reduced consumption of households (Coibion, Georgarakos, Kenny, et al. 2021).

One concern with these types of questions is that they might be cognitively demanding for many respondents. Probabilistic elicitations induce higher dropout rates from surveys, which might bias the inference one draws from the overall survey (D'Acunto et al. 2020). Even if providing an answer, respondents who do not

understand the question format might report a level of uncertainty that differs from the actual uncertainty in their beliefs. Moreover, the ordering of the inflation bins—for example, listing the inflation bins before the deflation bins—can prime respondents toward describing higher expected inflation. Also, these questions typically center around zero and have narrower bandwidths around zero than at the extreme ranges. These design features possibly induce survey participants to perceive that values close to zero are considered more likely by the designers of the survey. Furthermore, using a fixed set of bins for possible outcomes can be constraining in times of crisis or otherwise unusual times, so that responses can end up being lumped in extreme bins. This issue has no easy solution, because adjusting the size and values of bins across survey waves, as for example the Federal Reserve Bank of Philadelphia did with the Survey of Professional Forecasters in response to the Great Recession and the COVID-19 crisis, makes it difficult to compare survey responses over time. Finally, empirical research shows that survey participants might report distributions that feature holes, which likely reflects their inability to understand a probability distribution.

To address some of these issues, Delavande and Rohwedder (2008) propose simplified visual representations of probability masses that reduce the cognitive burden for respondents who have lower numerical literacy. Alternatively, Altig et al. (forthcoming) propose asking respondents to report five possible scenarios for a given variable and then assign probabilities for these scenarios.

Addressing the Challenges

Survey designers have been creative in addressing these potential issues. For example, the response rate for a survey run by a private firm is often 10 percent or less while government-run surveys have response rates of between 50 and 80 percent, so finding a way to rely on government survey tools is useful. Visual aids can help improve the response rates and quality of responses, especially for those who struggle with understanding questions or formulating responses (for example, Delavande et al. 2011). Quantitative questions can be complemented with easier-to-answer qualitative questions. Testing various elements of survey instruments can help quantify potential biases in responses. Generally, more educated, financially literate respondents (say, managers of firms) are less sensitive to variations in the wording of questions. Some forms of priming could be addressed fairly easily by methods like randomly changing the order of questions/options or making responses more open ended. Many of these challenges are directly tackled in more ad hoc surveys that researchers design for addressing specific research questions (D'Acunto, Malmendier, et al. 2021).

Systematic Patterns in the Inflation Expectations of Households and Firms

A set of facts about subjective inflation expectations that are common to both households and firms has been documented across surveys, elicitation methods,

locations, and time periods. Hence, these facts are not artifacts of the measurement challenges we discussed above. These common patterns include: i) a systematic upward bias in numerical inflation expectations when compared to both lagged realized inflation and the average numerical expectations of professional forecasters; ii) a large amount of disagreement about future inflation, including fat tails; iii) high uncertainty in forecasts of future inflation; iv) strong correlation between the updating of expectations at the individual level in short-run and long-run inflation forecasts; and v) predictability of inflation forecasts using perceived inflation. These facts have been detected for both households and firms, even though they are more pronounced among households (Link et al. 2021). Documenting these facts and their robustness across data sets, countries, and time periods helps to guide the empirical search for the determinants of household and firms' inflation expectations and how these expectations determine real decisions, which we discuss in the following sections.

Systematic Upward Bias in Inflation Expectations

Across different data sets, countries, and time periods, researchers have documented that the average and median numerical inflation expectations of households and firms tend to be higher than the realized inflation rates that occur subsequently, and also higher than the contemporaneous inflation expectations of professional forecasters and financial-market participants.

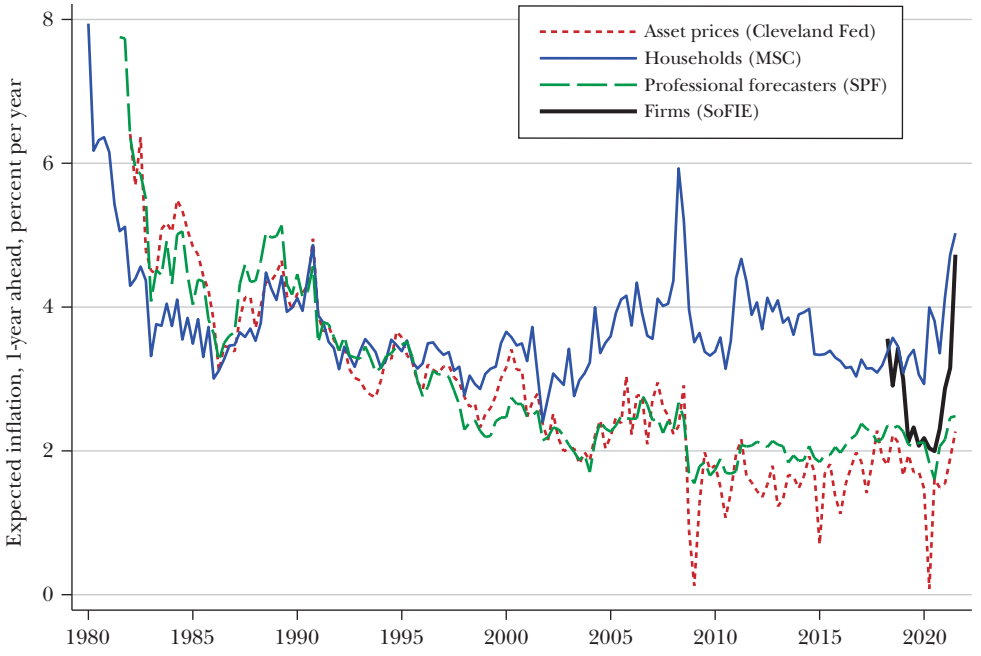
Figure 2 summarizes this pattern. The figure plots the mean of the numerical inflation expectations elicited from households each month in the Michigan Survey of Consumers as well as the mean response of top business executives participating in the Survey of Firms' Inflation Expectations.[3] For comparison, the figure also includes expert forecasts of inflation from the Survey of Professional Forecasters from the Federal Reserve Bank of Philadelphia and the time series of expectations extracted from asset prices by the Federal Reserve Bank of Cleveland using "inflation swaps" (a financial derivative in which investors "swap" a fixed payment for a set of payments based on the Consumer Price Index).

As the figure illustrates, the inflation expectations of US households were systematically higher than those of either professional forecasters or financial market participants over the last two decades. The inflation expectations of firms (at the far right-hand side of the figure) also depart significantly from those of experts, although the size of the upward bias varies more over time. Other work has documented the same patterns for households and firms in many other advanced economies characterized by low and stable inflation (for example, Candia et al. 2021b). The higher inflation expectations of households and firms is one of the most robust characteristics emanating from surveys of subjective expectations.

One way to gauge information about the source of such upward bias is assessing whether the bias is systematically larger or smaller for certain demographic groups,

[3]For more details see http://firm-expectations.org/.

Figure 2
Mean Inflation Expectations



Source: Candia et al. (2021a).
Note: Financial markets' expectations are from the Federal Reserve Bank of Cleveland, households' expectations are from the Michigan Survey of Consumers (MSC), professional forecasters' expectations are from the Survey of Professional Forecasters run by the Federal Reserve Bank of Philadelphia, managers' expectations are from the Survey of Firms' Inflation Expectations (SoFIE). We exclude responses of households that are greater than 15 percent or less than -2 percent. Firms' expectations are from our new survey of CEOs. We exclude responses that are greater than 15 percentage points or less than -2 percentage points. All moments are computed using survey weights.

which could point toward potential determinants of the bias. Indeed, the bias varies systematically across specific demographic groups: for example, the upward bias is systematically higher for women than for men (Bruine de Bruin et al. 2010; D'Acunto, Malmendier, and Weber 2021), a point to which we will return. Moreover, the bias is lower for agents who have higher cognitive abilities (D'Acunto et al. 2019). Also, socioeconomic status—a combination of formal education and income levels—helps to explain cross-sectional variation in several macroeconomic expectations (Das, Kuhnen, and Nagel 2020), including the size of the upward bias in inflation expectations (Bruine de Bruin et al. 2010; Angelico and Di Giacomo 2020; Weber, Gorodnichenko, and Coibion forthcoming): households from lower socioeconomic backgrounds tend to have systematically higher inflation expectations than others.

On the firm side, systematic differences in inflation expectations have been detected across industries (Coibion, Gorodnichenko, and Kumar 2018)—again, a

point to which we will return. The position of a respondent within a firm is also predictive of their inflation expectations: chief executive officers and chief financial officers have lower inflation expectations than other managers, who in turn have lower inflation expectations than the average employee, even after controlling for differences in education and income (Savignac et al. 2021). This variation suggests that the hierarchical role of those who set prices and wages in firms can be important because their wage- and price-setting decisions depend on inflation expectations that are closer or further away from the expectations of experts.

High Disagreement about Future Inflation

Surveys of households and firms display substantial dispersion of inflation expectations even within the same survey waves (Mankiw, Reis, and Wolfers 2004). Figure 3 reports the distribution of numerical inflation expectations across all waves of the Michigan Survey of Consumers (panel A), the Survey of Firms' Inflation Expectations (panel B), and the Survey of Professional Forecasters (panel C) from 2018:II to 2021:III. For households and firms, reported inflation expectations cover an extremely wide range of values, whereas those of professional forecasters are tightly concentrated around the mean.[4] This profound disagreement about aggregate inflation expectations might appear surprising, because all agents are asked to report expectations about the same macroeconomic variable, rather than about a personal-outcome variable. Similar patterns hold across surveys in the United States and abroad, so specific survey design features are unlikely to be the driving force of such systematic disagreement. Instead, the data point toward two potential directions in terms of determinants of aggregate expectations: variation in the information sources different agents use to form their expectations and variation in economic beliefs driven by a different interpretation of the same economic shocks that all agents face.

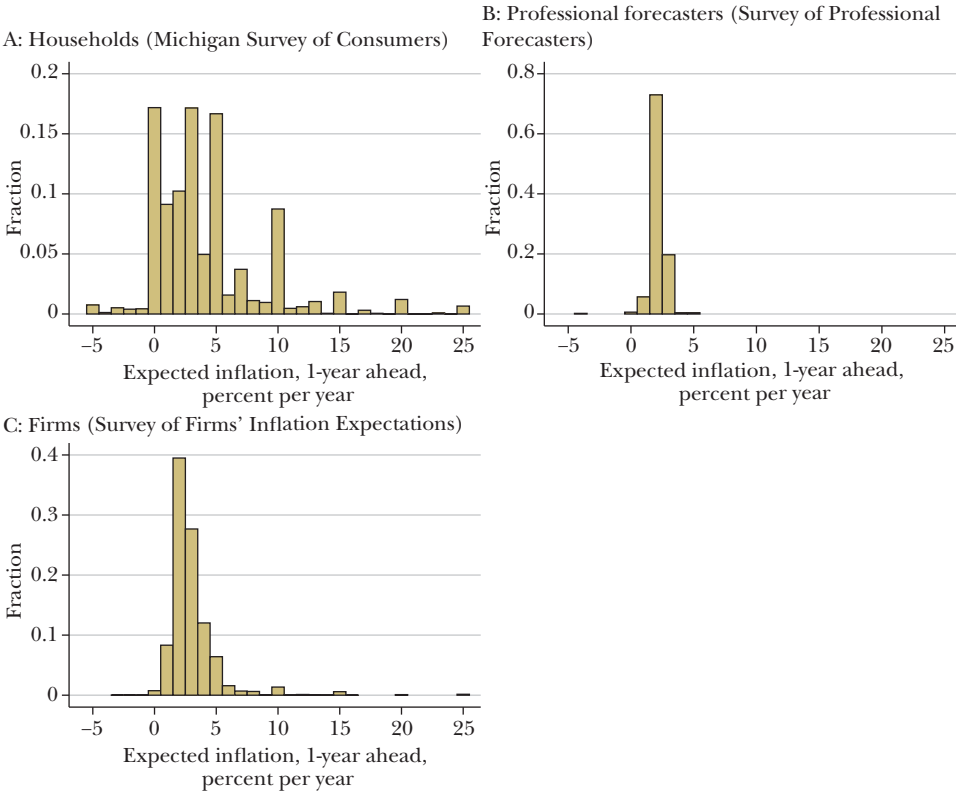
High Uncertainty in Inflation Expectations

There are several ways to gauge the level of uncertainty in inflation forecasts. Looking back at Figure 3, one feature is the extent to which households' and firms' expectations tend to be reported as multiples of five. This form of rounding has been interpreted as a proxy for respondents' uncertainty regarding the actual level of their inflation expectations (Binder 2017).

Another way to gauge the uncertainty in forecasts is having respondents assign probabilities to a range of possible outcomes for future inflation. Figure 4 presents results from doing so, focusing specifically on the probability that households in the Survey of Consumer Expectations from the Federal Reserve Bank of New York, firms in the Survey of Firms' Inflation Expectations, and professional forecasters in the Survey of Professional Forecasters assign to inflation being above either

[4]Professional forecasters might also have strategic considerations and might not want to deviate too much from an average forecast in either direction to avoid being perceived as overly pessimistic or optimistic and ultimately less credible.

Figure 3
Cross-sectional Dispersion in Expectations



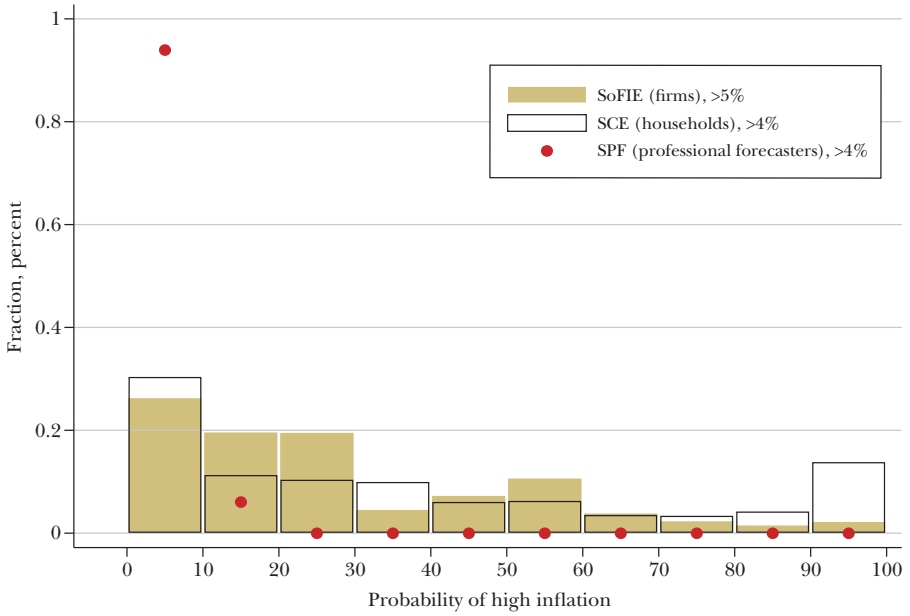
Source: Weber et al. (2022).
Note: The figure reports the distribution of short-term (1-year-ahead) inflation forecasts. Panel A shows results for households (Michigan Survey of Consumers). Panel B shows results for professional forecasters (Survey of Professional Forecasters). Panel C shows results for firms (Survey of Firms' Inflation Expectations). The distributions are computed using survey weights. The sample period covers 2018:II–2021:III.

4 or 5 percent in the next 12 months. For households and firms, these probabilities tend to be quite high, which indicates a wider range of uncertainty about the inflation outlook. For professional forecasters, the range of uncertainty is much lower. This relative difference in forecast confidence of professionals relative to households and firms has also been found to be a pervasive characteristic of inflation expectations for the general public.

Unanchored Inflation Expectations

We have so far restricted our attention to one-year ahead inflation forecasts, which is a relatively short horizon. Some surveys also ask respondents about inflation over longer time horizons, such as five or ten years. These longer-run expectations

Figure 4
Uncertainty in Inflation Expectations

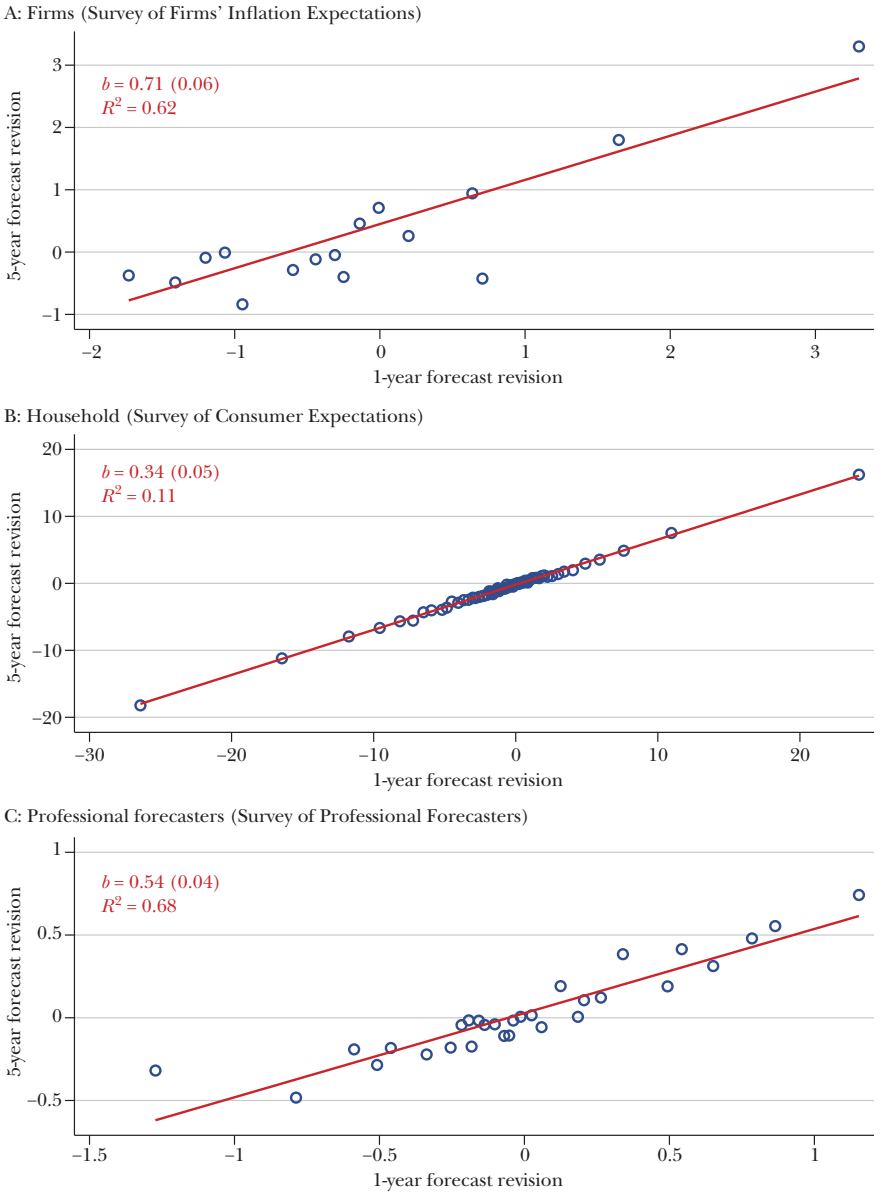


Source: Candia et al. (2021a).
Note: The histogram shows uncertainty for expected inflation in 2019:I. The Survey of Firms' Inflation Expectations (SoFIE) reports the distribution of the probability that inflation over the next 12 months will exceed 5 percent while the Survey of Consumer Expectations (SCE) and the Survey of Professional Forecasters (SPF) report the distribution of the probability that inflation over the next 12 months will exceed 4 percent. The distributions are computed using survey weights.

of inflation can be informative about the degree to which inflation expectations are anchored. Indeed, a common definition of "anchored" expectations is that changes in short-run inflation expectations should be largely uncorrelated with changes in long-run expectations: if one believes that the central bank is going to be successful in achieving its ongoing target for low inflation in the medium run, then current shocks to inflation should be offset by the central bank. For example, individuals trusting the central bank should expect tight monetary policy following inflationary shocks, and long-run expectations should therefore be insensitive to short-run fluctuations.

Figure 5 presents tests of this notion for households, firms, and professional forecasters by plotting the association between changes in individuals' one-year ahead inflation expectations across two adjacent survey waves with the change in their expectations about longer-run inflation. Strikingly, there is a strong positive correlation between these revisions, indicating that inflation expectations are *not* well anchored during this period from approximately 2017–2020. Shocks to the economy that lead individuals to expect higher inflation over the next year also

Figure 5
Correlation in Short-Term and Long-Term Inflation Expectations



Source: Weber et al. (2022).
Note: The binscatters show the relationship between 1-year-ahead and 5-year-ahead inflation forecasts. The sample period covers waves 2018:IV, 2019:IV, and 2020:IV for the Survey of Firms' Inflation Expectations, 2017:I–2020:IV for the Survey of Consumer Expectations, and 2018:I–2021:III for the Survey of Professional Forecasters. Huber robust regression is used to downweigh the importance of outliers and influential observations. Robust standard errors are in parentheses.

lead those individuals to expect significantly higher inflation over the next five to ten years, indicating that people do not think that inflation shocks are short-lived or that the central banks will take actions that offset these shocks.

Perceived and Expected Inflation

Information about aggregate inflation statistics is publicly available and regularly displayed on media outlets, so one might think that most individuals are aware of it, and yet we saw substantial amounts of disagreement across individuals and firms and large degrees of uncertainty. In fact, it turns out that what people believe about recent inflation is one of the strongest predictors of what they expect about future inflation. This result was first documented for Swedish households in Jonung (1981) and has repeatedly been verified since. Figure 6 plots this result for US households and firms in panels A and B: those who think that inflation has recently been high tend to be the same people as those who believe that future inflation will be high. The association is instead substantially weaker for professional forecasters in panel C. This evidence suggests that we can explain much of the variation in people's beliefs about the future through their perceptions about the recent past. We mentioned earlier that individuals might disagree either because of different opinions about how the economy works, leading them to anticipate a different evolution of prices in the future given the current state of the economy (Andre et al. 2021), or because they hold different views about the current state of the economy.

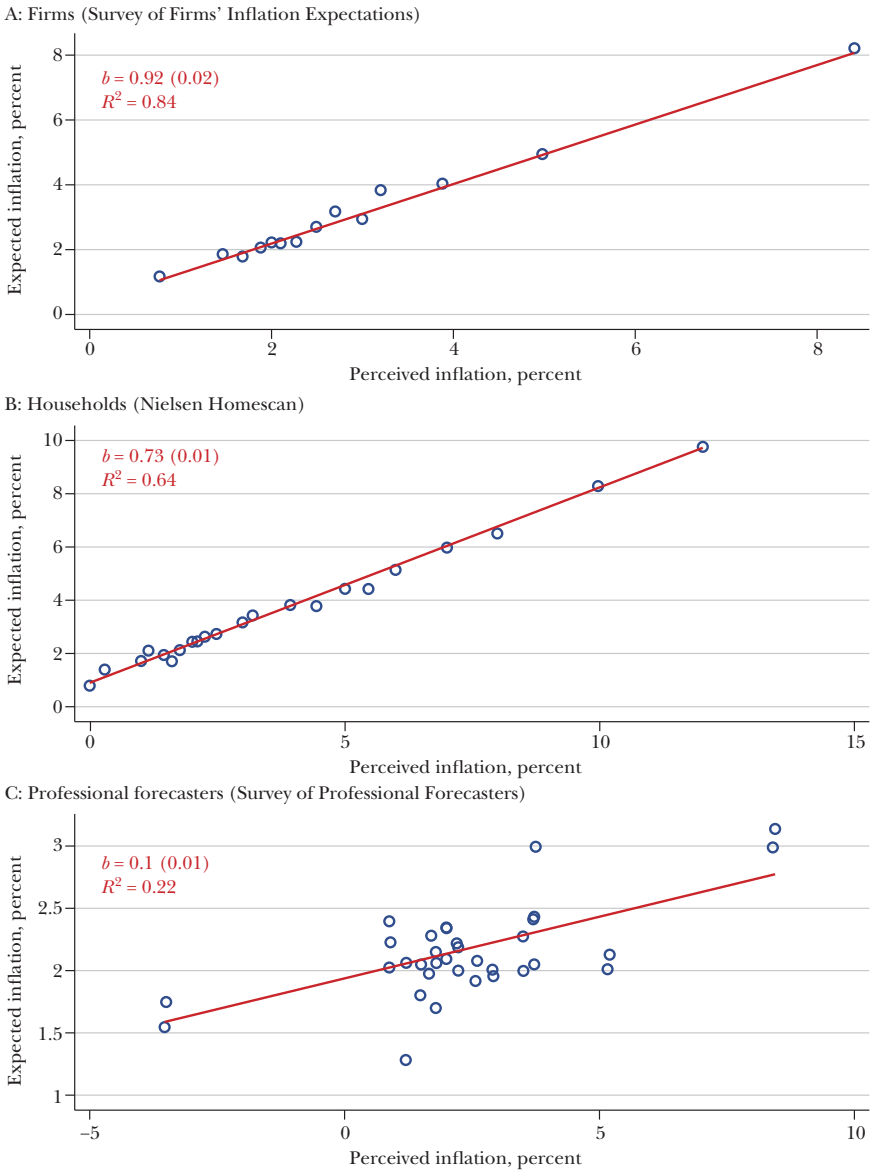
Determinants of Inflation Perceptions and Expectations for Households and Firms

If much of the differences in people's forecasts of future inflation stem from their different views about recent inflation dynamics, where does the disagreement about recent inflation dynamic arise? When households or business executives are asked about how they receive information about inflation, most report that their main source of information is their own shopping experience (D'Acunto, Malmendier, et al. 2021; Cavallo, Cruces, and Perez-Truglia 2017; Kumar et al. 2015), as well as family and friends. Another source that they emphasize is news and social media. In this section, we review existing evidence about the role these channels play in explaining underlying differences in perceived and expected inflation. We also discuss additional mechanisms that have been documented by recent research, including cognitive constraints and differences in incentives to pay attention to inflation. This research offers empirical guidance to macroeconomic theory as it seeks to understand how to model heterogeneous agents who form different expectations and hence make different economic choices.

Exposure to Heterogeneous Price Signals

Recent research on understanding inflation expectations has focused on the fact that even similar households and managers differ in the price signals they

Figure 6
Inflation Expectations and Perceptions



Source: Weber et al. (2022).
Note: The binscatters show the relationship between 1-year-ahead inflation forecasts and perceptions (nowcasts) of inflations. The sample period is 2018–2021. Panel A shows results for the Survey of Firms' Inflation Expectations. Panel B shows results for survey responses of households participating in the Nielsen HomeScan Panel; see Coibion, Gorodnichenko, and Weber (2022) for more details. Panel C shows results for the Survey of Professional Forecasters. Perceptions in the Survey of Professional Forecasters are measured as the nowcast for the most recent (or current) quarter-on-quarter annualized inflation rate. Perceptions and expectations for managers and households are restricted to [-2%,15%] range. Huber robust regression is used to downweigh the importance of outliers and influential observations. Robust standard errors are in parentheses.

observe in their environment and through daily activities, such as shopping for groceries or buying gasoline. Even if grocery bundles represent a relatively small fraction of the overall consumption basket of households, grocery price changes are quite salient, visible, and varied direct signals about price changes to which anybody who shops is exposed (D'Acunto, Malmendier, et al. 2021). Leveraging data from the Nielsen Consumer Panel for about 60,000 Americans, one can observe the nondurable goods individuals purchase and the exact prices they pay at the weekly frequency, due to the fact that these households use optical scanners to track all of their purchases. Customized surveys on this panel find that households who have observed the highest inflation rates in their own consumption bundles in the recent past have significantly higher expectations for general inflation over the following 12 months (see also Figure 1). This result holds for both point estimates and probability-distribution implied means as well as across elicitation methods, such as those in the Michigan Survey of Consumers and the New York Fed Survey of Consumer Expectations. This link is driven by the price changes of the goods that are purchased most frequently by each household, rather than by the expenditure share of goods in households' consumption bundles: someone who purchases milk frequently tends to think aggregate inflation is rising when they observe an increase in the price of the milk they purchase (D'Acunto, Malmendier, et al. 2021). Moreover, individuals tend to put a higher weight on positive prices changes than negative price changes, which helps to explain the general upward bias in expected inflation. In addition, initial price pressures in narrow categories of goods that are very salient to households can result in an immediate uptick of overall inflation expectations. We observed this pattern in April 2020 when the inflation expectations of households jumped upward following an increase in grocery prices, and again in the summer of 2021 when the price of used cars skyrocketed. Both times, economic forecasters as well as the Federal Reserve predicted low inflation or only temporary price pressures in narrow categories.

Observed price changes differ across individuals who have different grocery bundles as well as across individuals who shop at different outlets (Kaplan and Schulhofer-Wohl 2017; Weber, Gorodnichenko, and Coibion forthcoming). When asked about which price signals they consider when forming inflation expectations, women tend to mention the price of milk or bread, whereas men are more likely to mention the price changes of beer and gasoline (D'Acunto, Malmendier, and Weber 2021). The amount of shopping that agents do is also important. Individuals who report doing most of the shopping for their household typically have higher inflation expectations than those who do not. Women are more likely to be the primary shopper within their household, and the difference in the average inflation expectations of men and women previously mentioned disappears once one controls for who is primarily responsible for the shopping. Indeed, men who do the shopping in their household have the same average expectations of inflation as women who do the shopping, and the same is true for men and women who are not responsible for doing the shopping for their household (D'Acunto, Malmendier, and Weber 2021).

Another dimension that might bias inflation expectations is agents' limited memory of past prices. Individuals on average are correctly informed about the current price level, but they think that prices were cheaper in the past than what they actually were; they have a downward-biased memory of past prices (D'Acunto and Weber 2021). As a result of this bias in memory, perceptions of inflation arising from shopping will tend to be biased upward (Bordalo, Gennaioli, and Shleifer 2020; Enke, Schwerter, and Zimmermann 2020). These biases are also likely to be more persistent in agents' minds in times of major shocks to their environment and the set of price signals agents observe around them (Goldfayn-Frank and Wohlfart 2020).

Observed price signals influence aggregate inflation expectations not just of households, but also of firm managers, who focus on the price signals that they observe in their industries. For example, firms in sectors that have witnessed higher inflation recently tend to form higher beliefs about aggregate inflation, even when those industry-level price changes are unrelated to aggregate price changes (Andrade et al. forthcoming). The importance of directly observed price changes as an individual-level signal that helps to explain aggregate inflation expectations is a pervasive finding in the literature.

Similarly, the average inflation expectations of US households are particularly sensitive to changes in oil prices over time, which are the main determinant of the gasoline prices that are omnipresent in American life and one of the most frequently purchased items.

Media and Policy Communication

The fact that inflation expectations are on average biased upward for households and firms and dispersed across survey respondents suggests that households (and firms) might not devote much attention to media coverage of inflation or to public announcements, like press releases by the Federal Reserve—at least in low-inflation environments. Carroll (2003) estimates a model in which individuals update their expectations probabilistically based on news coverage of inflation and finds that individuals, on average, update their inflation expectations about once a year.

Another reason individuals might not devote much attention to media coverage of inflation and monetary policy is its complexity. For instance, households did not update their inflation expectations upward to the first forward guidance announcements by the European Central Bank (as theory would suggest) but instead adjusted inflation expectations sharply upward to announcements of future increases in consumption taxes (D'Acunto, Hoang, and Weber 2021; Bachmann et al. 2021).

The salience of policy in media and its complexity play a major role in how individuals set expectations. For example, announcement of changes in consumption taxes are discussed heavily not only by specialized media but also by popular media in print and online, whereas discussions of forward guidance by a central bank are more technical and tend to be relegated to specialized media sources, which most households do not consult. Policies like forward guidance are also more complex to

understand by ordinary households, because they require that agents understand that keeping interest rates low beyond the time it is warranted by future economic conditions will generate inflation in the future, and hence they should increase inflation expectations today. In contrast, announcing higher consumption taxes in the future directly tells households that prices will rise.

Households seem unaware of the dramatic policy announcements in recent decades. In August 2020, the Federal Reserve announced that monetary policy would shift from inflation targeting to "average inflation targeting"—so that if inflation was below its target for a time, the Fed would allow inflation to be above its target for an offsetting period in the future. However, the vast majority of US households heard no news about monetary policy in the days surrounding the announcement (Coibion et al. 2020a). Moreover, those who reported having heard news were not more likely to pick the correct policy framework in a multiple choice question and their inflation expectations did not differ from the expectations of individuals who reported not having heard any news. Focusing on more standard monetary policy news, Lamla and Vinogradov (2019) show in daily event studies around announcements by the Federal Open Market Committee that announcements do not affect households' subjective inflation expectations.

In short, the current conduct of monetary policy communication by the Federal Reserve and other central banks around the globe is likely ineffective in reaching ordinary households, contrary to more innovative forms of engagement such as the occasional use of reggae songs by the Central Bank of Jamaica or the use of Twitter as a communication tool by Olli Rehn, the Governor of the Bank of Finland (D'Acunto et al. 2020).[5]

To study the potential role of communication on the inflation expectations of households and firms in case central banks were able to reach them with their communication, a growing body of work uses information provision experiments within surveys. In fact, a stated goal of leading central banks is being heard and understood by ordinary people. Christine Lagarde (2020), president of the European Central Bank, stressed the importance of the audience at a hearing in front of the European Parliament when she said: "After all, it is the everyday economic decisions of people and companies that we seek to influence with our policy and communication. If our language is not accessible, our policy will be less effective." A typical paper in this literature elicits inflation expectations, and then randomly splits the sample of survey participants into treatment and control groups, provides different information like inflation forecasts or inflation targets to individuals in the treatment groups, and elicits posterior inflation expectations identically for all survey participants. The updating of expectations relative to the survey participants in the control condition thus provides the causal treatment effect of the provided

[5]More generally, policy communication should be more accessible to the general public by making messages easier to understand (for example, Bholat et al. 2019; Haldane and McMahon 2018). Blinder et al. (2008) provide an early survey of the literature on the importance of policy communication for monetary policy.

information on inflation expectations. Providing information about simple summary statistics of inflation such as current, past, or expected inflation and the Fed inflation target results in large average revisions of inflation expectations in the range of 1 to 1.5 percentage points (Coibion, Gorodnichenko, and Weber 2022; Coibion, Georgarakos, Gorodnichenko et al. 2021). Providing individuals with the full Federal Open Market Committee press release, which contains these statistics, but also more technical details and context, results in an average forecast revision of similar magnitude. However, the survey participants who instead received the coverage of the Federal Open Market Committee announcement from a media source (in this case, *USA Today*) revised their expectations by less than half of the revisions of other survey participants. The need to read a text of several paragraphs and comprehend its content cannot explain this difference, because the Fed announcement includes more jargon and complexity than the media article. (A possible lack of credibility of *USA Today* relative to other newspapers is also an unlikely explanation, because *USA Today* ranks higher in terms of credibility for economics and business than the *New York Times*, *Wall Street Journal*, or *Washington Post*.) Instead, traditional news media have low credibility and attract lower trust than other sources in a representative sample of 25,000 Americans. In particular, survey participants with low income and low formal education barely react to the media treatment, whereas they do react to the Fed statement.

Overall, the muted impact of official releases, communication, and the media on inflation expectations is consistent with individuals reporting that they predominantly rely on the price changes they observe in their own shopping when forming inflation expectations—in line with the famous Lucas (1972) "islands" model.

Cognitive Constraints

In addition to the large differences in perceived inflation due to different exposure to price signals in daily life, heterogeneous cognitive abilities contribute to shape inflation expectations. Nordic countries like Finland allow the linking of measures of cognitive abilities for all men—IQ as measured by a military entrance test—at the individual level together with survey data on inflation expectations and consumption plans. Individuals at the bottom of the IQ distribution display absolute forecast errors for inflation that are larger by a factor of two relative to those at the top of the distribution. Forecast errors decline monotonically as IQ rises, and hence this systematic difference is not driven by either individuals with the lowest or highest cognitive abilities (D'Acunto et al. 2019, forthcoming). Relating consumption plans to inflation expectations reveals that only men above the median level of IQ increase their planned spending when they expect higher inflation, as intertemporal substitution would predict. Differences in financial constraints, formal education, or income, by contrast, do not matter for these associations after controlling for IQ.

D'Acunto, Hoang, et al. (2019, 2021, forthcoming) also find that respondents with different levels of cognitive abilities think about substantially different concepts of inflation when answering surveys: low-IQ respondents predominantly think about the price changes of a few concrete goods they have in mind, whereas

high-IQ respondents are more likely to think about the abstract concept of inflation and its relation with other macroeconomic variables. Moreover, low-IQ respondents think that high inflation tends to be associated with bad economic times and that persistent deflation is desirable, which helps explain why they do not increase consumption when they expect higher inflation.

Taken together, these results suggest that differences in cognitive abilities play an important role in shaping inflation expectations and help inform recent advances in macroeconomic theory on how to model heterogeneous agents and agents with limited cognition for the transmission of fiscal and monetary policy (Woodford 2019; Farhi and Werning 2019).

Incentives to Gather Information about Inflation

Some households and firms have a greater perceived return to gather information about inflation. We have already discussed one example of this incentive effect: business executives and managers tend to know more about average inflation than households, but less than professional forecasters. Moreover, households with higher incomes or who own mortgages tend to have more accurate inflation expectations. Another aspect that drives the incentive to be informed is the level and volatility of realized inflation. Households in high inflation countries tend to also be more informed about inflation (Cavallo, Cruces, and Perez-Truglia 2017). Many more examples of incentive effects have been documented in the literature.

In the case of firms, one key determinant of managers' informedness about inflation is the number of competitors their firms face. A survey of firms in New Zealand revealed that as firms face more competitors, their knowledge of inflation dynamics increases (Afrouzi 2020; Coibion, Gorodnichenko, and Kumar 2018). Firms that sell a larger and wider range of products pay more attention to aggregate inflation increases, whereas firms that sell a more limited number of products find it sufficient to be informed about prices in their own specific market (Yang 2020). Also, firms anticipating changing prices in the near future acquire more information about inflation to guide their pricing decisions, whereas firms not expecting to change prices for many months are less well-informed.

Inflation Expectations and Economic Choices

Based on standard macroeconomic and intertemporal microeconomic models, the extent to which households and firms expect prices to rise should matter for many decisions—saving and consumption choices, wage bargaining and labor supply, as well as investment, leverage, hiring, and price-setting decisions. Seeking causal evidence about whether inflation expectations do actually affect decisions has become an active area of research in recent years, because if agents use their inflation expectations when making choices, the heterogeneity in choices we observe in the data might be explained by the same determinants as those of inflation expectations.

Inflation Expectations and Household Choices

Intuitively, when households anticipate higher price growth in the future, they should choose to consume more today before those price increases materialize. Spending on durable goods should be affected most, because they are easier to substitute intertemporally than non-durable goods.

This theoretical prediction was first explored at the individual level in Bachmann et al. (2015) using data from the Michigan Survey of Consumers. On average, they found no correlation between individuals' willingness to purchase large-ticket items and their inflation expectations, although a positive correlation was detected among highly educated respondents and those who had inflation expectations close to the subsequent realization of inflation. However, this survey is largely a cross-sectional dataset—that is, the same individuals are not tracked for extended periods of time—and large dispersion in inflation expectations might complicate the empirical analysis if differences in average expected inflation rates exist across individuals. Indeed, focusing on changes in inflation expectations within individuals over time, Vellekoop and Wiederholt (2020) document a positive association between inflation expectations and consumption choices. Using data from Finland, D'Acunto et al. (forthcoming) document facts that reconcile these results in the literature. First, they find that controlling for heterogeneous characteristics is central to establishing a positive association between inflation expectations and the willingness to purchase durable goods. Second, cognitive abilities shape the strength of this association between inflation expectations and consumption decisions. This result is independent of financial constraints, formal education, or other observable characteristics and could be interpreted as a "human friction" (D'Acunto, Hoang, et al. 2021), which limits the transmission of economic policy interventions that operate through households' inflation expectations.

Recent research has used randomized control trials to identify how expectations shape decisions. Researchers randomly allocate survey participants to different groups: some groups receive information about inflation or monetary policy (the "treatment" groups) while others do not (the "control" group). By comparing the inflation expectations of the individuals who received information to the control group, researchers can determine how information changes expectations. In some cases, the background information that alters beliefs of one group in a survey about future price increases can also arise from a natural experiment, as in the case of a pre-announced increase in consumption taxes (D'Acunto, Hoang, and Weber 2022). Following the announcement of higher future consumption taxes, most individuals who expect higher inflation going forward, relative to their baseline assessment of price changes, report that now is a good time to consume and especially to purchase durable goods.

A related approach uses randomized control trials not just to create exogenous variation in inflation expectations but also to study how these changes in inflation expectations affect subsequent consumption decisions (for a recent review, see Haaland, Roth, and Wohlfart forthcoming). Coibion, Gorodnichenko, and Weber

(forthcoming) use customized surveys on the Nielsen HomeScan Panel and find that, in both survey and actual scanner data, households with higher expected inflation increase their nondurable consumption for up to six months after the survey intervention. Because the Nielsen HomeScan Panel does not contain large-ticket items, they focus on surveys three and six months after the experimental variation to study whether higher inflation expectations induced individuals to change their purchases of durable goods. Contrary to economic theory, they find that higher inflation expectations result in a lower likelihood that individuals purchased larger-ticket items in the months after treatments. Other studies also using experimental variation find similar results in the United States and the Netherlands (Coibion et al. 2019; Coibion, Georgarakos, Gorodnichenko, et al. 2021), which might be driven by individuals associating higher inflation with worse economic outcomes (Andre et al. 2021; Kamdar 2019; D'Acunto et al. forthcoming). Subjective uncertainty about inflation is also important to explain saving choices—more uncertain households, even controlling for demographic characteristics, display more precautionary consumption, credit, and investment decisions (Fermand et al. 2018).

This evidence suggests that households do in fact use their inflation expectations when making economic decisions. But the inconsistent evidence across studies and across types of goods indicates that the literature has not yet fully grasped the mechanisms and models households use when relating inflation expectations to consumption decisions.

Besides consumption and savings choices, inflation expectations should also influence individual decisions about borrowing, including their mortgage choices (Botsch and Malmendier 2020), as well as their wage bargaining and labor supply decisions. So far, systematic evidence for these outcomes is limited, in part because of a lack of viable data. Research can make progress on these questions using customized survey data linking expectations with actual decisions.

Inflation Expectations and Firms' Choices

The decisions that firms make about price-setting, labor demand, investment, and leverage directly depend on their inflation expectations. Two recent studies provide causal evidence from randomized control trials that changes in inflation expectations shape firms' decisions: one from New Zealand (Coibion et al. 2018) and one from Italy (Coibion et al. 2019). In each country, a subset of firms was provided with information about inflation or monetary policy, while a control group received no such information. The information had pronounced effects on the inflation expectations of the treated firms. These two studies then tracked the decisions of firms over time to discern if and how changes in beliefs changed their economic decisions. While conceptually similar, the two studies differed in the countries considered, the duration of the information treatments (one-time in New Zealand versus repeated over years in Italy), the monetary policy regime (Italy was at the zero interest rate lower-bound for part of the sample), how outcomes were measured (self-reported actions in New Zealand versus administrative data in Italy), and the types of firms (the New Zealand study had primarily small firms

while the Italian study had primarily larger firms). Despite these differences, both studies found decisive evidence that changes in inflation expectations, induced by randomly allocated information treatments, had pronounced effects on the economic decisions of firms. Employment and investment decisions were found to be particularly sensitive to inflation expectations, while prices were only found to respond in Italy.

Firms' price-setting decisions also directly affect overall inflation. However, little research exists for how firms' pass-through of marginal costs of inputs into their prices depends on their expectations of future inflation.

Conclusion

Inflation expectations affect the economic decisions of both households and firms and for this reason have been thrust into the limelight by policymakers for decades. Academic research has been making progress in documenting and understanding how firms and households form their beliefs about future inflation and how these beliefs feed into the economic decisions of both households and firms. Research so far has also shown that heterogeneity in the determinants of inflation expectations can help make sense of the heterogeneous economic choices of otherwise similar households and firms as well as heterogeneous reactions to the same economic shocks by different households and firms.

For central banks, inflation expectations have become a key part of the conduct of monetary policy. The Federal Reserve, for instance, has often relied on relatively stable long-run inflation expectations to assess policy choices. As Jerome Powell (2020) said, "[E]xpected inflation feeds directly into the general level of interest rates. Well-anchored inflation expectations are critical for giving the Fed the latitude to support employment when necessary without destabilizing inflation." In theory, it would even be possible for a central bank to encourage higher inflation expectations as a form of monetary stimulus, since those who expect higher inflation in the future will perceive a correspondingly lower real interest rate in the present. However, central banks that want to manage inflation expectations as a policy tool have to be cautious, because raising inflation expectations could in fact backfire if households associate higher inflation with worse economic times (Coibion et al. 2020b).

The extent to which long-run inflation expectations are anchored, and the extent to which they will remain anchored, has played an important role in monetary decision-making in 2022 in response to the surge of inflation that began in 2021. However, the ability of policymakers to shape inflation expectations is under-studied and remains a point of contention in the literature. For example, households have been shown to understand simple messages about the aims of monetary policy interventions: a common example is the "whatever it takes" speech by former European Central Bank president Mario Draghi (2012), which conveyed the commitment of the central bank to supply as much liquidity as needed in troubled markets. On the

other side, households often barely react at all to communication about monetary policy instruments such as large-scale asset purchases (D'Acunto et al. 2020). The identity of the sender matters too. D'Acunto, Fuster, and Weber (2021) show that groups that have been historically underrepresented on monetary policymaking bodies, such as women and minorities, are more likely to form expectations in line with provided official forecasts when the forecasts are associated with a female or Black policymaker. These challenges highlight that the current conduct of monetary policy communication often does not reach ordinary households and firms and calls for more innovative communication tools.

The rebound of inflation levels around the world has turned the research questions of the evolution and management of inflation expectations into urgent policy questions, too. A detailed map of the effects of inflation expectations on multiple economic choices is crucial to assess the potential role of expectations as a monetary policy tool.

References

Afrouzi, Hassan. 2020. "Strategic Inattention, Inflation Dynamics, and the Non-Neutrality of Money." CESifo Working Paper 8218.

Altig, David, Jose Barrero, Nicholas Bloom, Steven Davis, Brent Meyer, and Nicholas Parker. Forthcoming. "Surveying Business Uncertainty." *Journal of Econometrics***.**

Andrade, Philippe, Olivier Coibion, Erwan Gautier, and Yuriy Gorodnichenko. Forthcoming. "No Firm Is an Island? How Industry Conditions Shape Firms' Expectations." *Journal of Monetary Economics***.**

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart. 2021. "Subjective Models of the Macroeconomy: Evidence from Experts and a Representative Sample." Unpublished.

Ang, Andrew, Geert Bekaert, and Min Wei. 2007. "Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?" *Journal of Monetary Economics* 54 (4): 1163–212.

Angelico, Cristina, and Federica Di Giacomo. 2020. "Heterogeneity in Inflation Expectations and Personal Experience." Unpublished.

Armantier, Olivier, Wändi Bruine de Bruin, Simon Potter, Giorgio Topa, Wilbert van der Klaauw, and Basit Zafar. 2013. "Measuring Inflation Expectations." *Annual Review of Economics* 5 (1): 273–301.

Bachmann, Rüdiger, Tim O. Berg, and Eric R. Sims. 2015. "Inflation Expectations and Readiness to Spend: Cross-Sectional Evidence." *American Economic Journal: Economic Policy* 7 (1): 1–35.

Bachmann, Ruediger, Benjamin Born, Olga Goldfayn-Frank, Georgi Kocharkov, Ralph Luetticke, and Michael Weber. 2021. "A Temporary VAT Cut as Unconventional Fiscal Policy." NBER Working Paper 29442.

Bernanke, Ben S., and Kenneth N. Kuttner. 2005. "What Explains the Stock Market's Reaction to Federal Reserve Policy?" *Journal of Finance* 60 (3): 1221–57.

Bholat, David, Nida Broughton, Janna Ter Meer, and Eryk Walczak. 2019. "Enhancing Central Bank Communications Using Simple and Relatable Information." *Journal of Monetary Economics* 108 (C): 1–15.

Binder, Carola C. 2017. "Measuring Uncertainty Based on Rounding: New Method and Application to Inflation Expectations." *Journal of Monetary Economics* 90 (C): 1–12.

Blinder, Alan S., Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen. 2008. "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence." *Journal of Economic Literature* 46 (4): 910–45.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2020. "Memory, Attention, and Choice." *Quarterly Journal of Economics* 135 (3): 1399–442.

Botsch, Matthew J., and Ulrike Malmendier. 2020. "The Long Shadows of the Great Inflation: Evidence from Residential Mortgages." CEPR Discussion Paper 14934.

Bruine de Bruin, Wändi, Wilbert van der Klaauw, Julie S. Downs, Baruch Fischhoff, Giorgio Topa, and Olivier Armantier. 2010. "Expectations of Inflation: The Role of Demographic Variables, Expectation Formation, and Financial Literacy." *Journal of Consumer Affairs* 44 (2): 381–402.

Bruine de Bruin, Wändi, Wilbert van der Klaauw, Giorgio Topa, Julie S. Downs, Baruch Fischhoff, and Olivier Armantier. 2012. "The Effect of Question Wording on Consumers' Reported Inflation Expectations." *Journal of Economic Psychology* 33 (4): 749–57.

Bullard, James. 2016. "Inflation Expectations Are Important to Central Bankers, Too." *Regional Banker* April 13. https://www.stlouisfed.org/publications/regional-economist/april-2016/inflation-expectations-are-important-to-central-bankers-too.

Bureau of Labor Statistics (BLS). 2021. "Household and Establishment Survey Response Rates." United States Department of Labor. https://www.bls.gov/osmr/response-rates/home.htm (accessed September 15, 2021).

Candia, Bernardo, Olivier Coibion, and Yuriy Gorodnichenko. 2021a. "The Inflation Expectations of U.S. Firms: Evidence from a New Survey." NBER Working Paper 28836.

Candia, Bernardo, Olivier Coibion, and Yuriy Gorodnichenko. 2021b. "The Macroeconomic Expectations of Firms." Unpublished.

Carroll, Christopher D. 2003. "Macroeconomic Expectations of Households and Professional Forecasters." *Quarterly Journal of Economics* 118 (1): 269–98.

Cavallo, Alberto, Guillermo Cruces, and Ricardo Perez-Truglia. 2017. "Inflation Expectations, Learning, and Supermarket Prices: Evidence from Survey Experiments." *American Economic Journal: Macroeconomics* 9 (3): 1–35.

Coibion, Olivier, Erwan Gautier, Yuriy Gorodnichenko, and Frédérique Savignac. 2021. "Firms' Inflation Expectations: New Evidence from France." NBER Working Paper 29376.

Coibion, Olivier, Dimitris Georgarakos, Yuriy Gorodnichenko, and Maarten van Rooij. 2019. "How Does Consumption Respond to News about Inflation? Field Evidence from a Randomized Control Trial." NBER Working Paper 26106.

Coibion, Olivier, Dimitris Georgarakos, Yuriy Gorodnichenko, and Michael Weber. 2021. "Forward Guidance and Household Expectations." NBER Working Paper 26778.

Coibion, Olivier, Dimitris Georgarakos, Geoff Kenny, Yuriy Gorodnichenko, and Michael Weber. 2021. "The Effect of Macroeconomic Uncertainty on Household Spending." NBER Working Paper 28625.

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

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

Coibion, Olivier, and Yuriy Gorodnichenko. 2015b. "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review* 105 (8): 2644–78.

Coibion, Olivier, Yuriy Gorodnichenko, Edward S. Knotek II, and Raphael Schoenle. 2020a. "Average Inflation Targeting and Household Expectations." NBER Working Paper 27836.

Coibion, Olivier, Yuriy Gorodnichenko, and Saten Kumar. 2018. "How Do Firms Form Their Expectations? New Survey Evidence." American Economic Review 108 (9): 2671–713.

Coibion, Olivier, Yuriy Gorodnichenko, Saten Kumar, and Mathieu Pedemonte. 2020b. "Inflation Expectations as a Policy Tool?" *Journal of International Economics* 124 (C): 1–27.

Coibion, Olivier, Yuriy Gorodnichenko, and Tiziano Ropele. 2020. "Inflation Expectations and Firm Decisions: New Causal Evidence." *Quarterly Journal of Economics* 135 (1): 165–219.

Coibion, Olivier, Yuriy Gorodnichenko, and Michael Weber. 2022. "Monetary Policy Communications and their Effects on Household Inflation Expectations." *Journal of Political Economy*.

D'Acunto, Francesco, Andreas Fuster, and Michael Weber. 2021. "Diverse Policy Committees Can Reach Underrepresented Groups." Becker Friedman Institute Working Paper 2021–95.

D'Acunto, Francesco, Daniel Hoang, Maritta Paloviita, and Michael Weber. 2019. "Cognitive Abilities and Inflation Expectations." *AEA Papers and Proceedings* 109: 562–6.

D'Acunto, Francesco, Daniel Hoang, Maritta Paloviita, and Michael Weber. 2020. "Effective Policy

Communication: Targets versus Instruments." Becker Friedman Institute Working Paper 2020–148.

D'Acunto, Francesco, Daniel Hoang, Maritta Paloviita, and Michael Weber. 2021. "Human Frictions in the Transmission of Economic Policy." NBER Working Paper 29279.

D'Acunto, Francesco, Daniel Hoang, Maritta Paloviita, and Michael Weber. Forthcoming. "IQ, Expectations, and Choice." *Review of Economic Studies***.**

D'Acunto, Francesco, Daniel Hoang, and Michael Weber. 2022. "Managing Households' Expectations with Unconventional Policies." *Review of Financial Studies* 35 (4): 1597–642.

D'Acunto, Francesco, Ulrike Malmendier, Juan Ospina, and Michael Weber. 2021. "Exposure to Grocery Prices and Inflation Expectations." *Journal of Political Economy* 129 (5): 1615–39.

D'Acunto, Francesco, Ulrike Malmendier, and Michael Weber. 2021. "Gender Roles Produce Divergent Economic Expectations." *Proceedings of the National Academy of Sciences* 118 (21): 1–10.

D'Acunto, Francesco, and Alberto Rossi. 2021. "Robo-Advising." In *The Palgrave Handbook of Technological Finance*, edited by Raghavendra Rau, Robert Wardrop, and Luigi Zingales, 725–49. Cham: Palgrave Macmillan.

D'Acunto, Francesco, and Michael Weber. 2021. "Memory and Beliefs: Evidence from the Field." Unpublished.

Das, Sreyoshi, Camelia M. Kuhnen, and Stefan Nagel. 2020. "Socioeconomic Status and Macroeconomic Expectations." *Review of Financial Studies* 33 (1): 395–432.

Delavande, Adeline, Xavier Giné, and David McKenzie. 2011. "Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence." *Journal of Development Economics* 94 (2): 151–63.

Delavande, Adeline, and Susann Rohwedder. 2008. "Eliciting Subjective Probabilities in Internet Surveys." *Public Opinion Quarterly* 72 (5): 866–91.

Draghi, Mario. 2012. "Speech by Mario Draghi, President of the European Central Bank at the Global Investment Conference in London 26 July 2012." Speech, Global Investment Conference, London, July 26. https://www.ecb.europa.eu/press/key/date/2012/html/sp120726.en.html.

Enke, Benjamin, Frederik Schwerter, and Florian Zimmermann. 2020. "Associative Memory and Belief Formation." NBER Working Paper 26664.

Farhi, Emmanuel, and Iván Werning. 2019. "Monetary Policy, Bounded Rationality, and Incomplete Markets." *American Economic Review* 109 (11): 3887–928.

Fermand, Elyas, Camelia M. Kuhnen, Geng Li, and Itzhak Ben-David. 2018. "Expectations Uncertainty and Household Economic Behavior." NBER Working Paper 25336.

Gabaix, Xavier. 2020. "A Behavioral New Keynesian Model." *American Economic Review* 110 (8): 2271–327.

Goldfayn-Frank, Olga, and Johannes Wohlfart. 2020. "Expectation Formation in a New Environment: Evidence from the German Reunification." *Journal of Monetary Economics* 115 (C): 301–20.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. Forthcoming. "Designing Information Provision Experiments." *Journal of Economic Literature***.**

Haldane, Andrew, and Michael McMahon. 2018. "Central Bank Communications and the General Public." *AEA Papers and Proceedings* 108: 578–83.

Jonung, Lars. 1981. "Perceived and Expected Rates of Inflation in Sweden." *American Economic Review* 71 (5): 961–68.

Kamdar, Rupal. 2019. "The Inattentive Consumer: Sentiment and Expectations." Unpublished.

Kaplan, Greg, and Sam Schulhofer-Wohl. 2017. "Inflation at the Household Level." *Journal of Monetary Economics* 91 (C): 19–38.

Kim, GwangMin, and Carola Binder. 2020. "Learning-through-Survey in Inflation Expectations." Unpublished.

Kumar, Saten, Hassan Afrouzi, Olivier Coibion, and Yuriy Gorodnichenko. 2015. "Inflation Targeting Does Not Anchor Inflation Expectations: Evidence from Firms in New Zealand." *Brookings Papers on Economic Activity* 46 (2): 151–225.

Lagarde, Christine. 2020. "Introductory statement by Christine Lagarde, President of the ECB at the ECON Committee of the European Parliament." Hearing at the Committee on Economic and Monetary Affairs of the European Parliament, Brussels, February 6, 2020. https://www.ecb.europa.eu/press/key/date/2020/html/ecb.sp200206~edb83d06a3.en.html

Lamla, Michael J., and Dmitri V. Vinogradov. 2019. "Central Bank Announcements: Big News for Little People?" *Journal of Monetary Economics* 108: 21–38.

Link, Sebastian, Andreas Peichl, Christopher Roth, and Johannes Wohlfart. 2021. "Information Frictions among Firms and Households." Unpublished.

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

Mankiw, N. Gregory, and Ricardo Reis. 2002. "Sticky Information versus Sticky Prices: a Proposal to Replace the New Keynesian Phillips Curve." *Quarterly Journal of Economics* 117 (4): 1295–328.

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

Manski, Charles. 2004. "Measuring Expectations." *Econometrica* 72 (5): 1329–76.

Michigan Survey of Consumers. 2022. "Reports." University of Michigan. https://data.sca.isr.umich.edu/.

NielsenIQ. 2017. "Chicago Booth Expectations and Attitudes Survey." Nielsen and NielsenIQ Mark https://www.chicagobooth.edu/research/kilts/datasets/nielsenIQ-nielsen.

Powell, Jerome. 2020. "New Economic Challenges and the Fed's Monetary Policy Review." Speech, Navigating the Decade Ahead Symposium, Jackson Hole, Wyoming. August 27, 2020. https://www.federalreserve.gov/newsevents/speech/powell20200827a.htm.

Powell, Jerome. 2021. "Transcript of Chair Powell's Press Conference September 22, 2021." Press Conference, Federal Reserve, September 22, 2021. https://www.federalreserve.gov/mediacenter/files/FOMCpresconf20210922.pdf.

Savignac, Frédérique, Erwan Gautier, Yuriy Gorodnichenko, and Olivier Coibion. 2021. "Firms' Inflation Expectations: New Evidence from France." NBER Working Paper 29376.

Vellekoop, Nathanael, and Mirko Wiederholt. 2020. "Inflation Expectations and Choices of Households." Unpublished.

Weber, Michael, Francesco D'Acunto, Yuriy Gorodnichenko, and Olivier Coibion. 2022. "Replication data for: The Subjective Inflation Expectations of Households and Firms: Measurement, Determinants, and Implications." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E169781V1.

Weber, Michael, Yuriy Gorodnichenko, and Olivier Coibion. Forthcoming. "The Expected, Perceived, and Realized Inflation of U.S. Households before and during the COVID19 Pandemic." *IMF Economic Review*.

Woodford, Michael. 2003. "Imperfect Common Knowledge and the Effects of Monetary Policy." In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honour of Edmund S. Phelps*, edited by Philippe Aghion, Roman Frydman, Joseph Stiglitz, and Michael Woodford, 25–58. Princeton: Princeton University Press.

Woodford, Michael. 2019. "Monetary Policy Analysis When Planning Horizons Are finite." *NBER Macroeconomics Annual* 33 (1): 1–50.

Yang, Choongryul. 2020. "Rational Inattention, Menu Costs, and Multi-Product Firms: Micro Evidence and Aggregate Implications." Unpublished.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 185–210

Blending Theory and Data: A Space Odyssey

Dave Donaldson

What are the effects of trade liberalization, or the recent US-China trade war? Is urban gentrification leading to spatial inequalities and an erosion of opportunities for economic mobility? Do transportation infrastructure investments justify their astronomic price tags? These are all great questions—and they comprise only a small sampling from the bread-and-butter topics of spatial economics. But readers seeking specific answers to such questions have come to the wrong place. The focus of this article is instead about how economists working in the fields of international, regional, and urban economics arrive at answers to these sorts of counterfactual—that is, inherently causal—questions. It is about the spatial journey rather than the spatial destination.

For questions like these, economic theory alone does little to narrow the range of quantitative answers. Moreover, for questions like these, nature has not granted us sufficiently rich quasi-experimental serendipity that we can draw on it as a replacement for economic theory. What is to be done? The only option on the table is to combine the lessons of economic theory with what we can glean from empirical patterns. While there are many ways to do so, my focus will be on research that pursues an explicit theory-empirics nexus. This process involves using the relationships identified by quasi-experimental variation to the full extent possible, while also recognizing that data typically will not fully answer the policy question that motivates a given research study. As such, the analyst must use additional

■ *Dave Donaldson is Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is ddonald@mit.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.185.

information—modeling assumptions and the logical deductions that follow from them—to bridge the gap between what is identified and what is desired.

Just as any theoretical model is a metaphor, not an attempt to be a true representation of reality, the work I describe in this article involves researchers aiming to build an *empirical* metaphor. Economists are used to resisting the temptation to judge a model by the strength of its abstraction or its assumptions. Instead, we ask how useful a model appears to be at achieving some goal. That is, the role of a model is to provide a clear mapping from assumptions to answers to a given question, and it should be judged relative to how faithfully it delivers that answer. An empirical model is no different. It provides a clear mapping of assumptions to answers—but it does so conditional on the extra information provided by features that can be observed in the data. In this sense, a theme that appears throughout much of my discussion is one in which researchers, aiming to answer a given question, understand that theoretical assumptions will be needed to answer their question, but still do their best to minimize the need for such assumptions through the use of facts that can be extracted from the available data.

The goal of this article is to highlight, through a generic framework and a range of examples, some of the techniques deployed in spatial economics that have leaned on the complementarity between theory and data. This inevitably draws on advances made in all areas of economics, and hence relates to other recent methodological surveys that emphasize interactions between theory and data.[1] Nevertheless, the nature of spatial research often presents unique challenges due to the large number of economic interactions at work both within geographic units (among producers, consumers, and factors of production) and across them.

Models and Questions

I begin by describing a generic research problem—a question to be answered, a set of data features that are observable, a set of beliefs about sources of exogenous variation in such data, and a "model."

My discussion revolves around the following scenario. We imagine a researcher who, in some setting of interest, desires to answer the question, "What would be the change in outcome W if a change in policy X were to occur?" Notably, the goal is to quantify a causal effect: that of X on W. To continue an example from above, with knowledge of the effect of certain transportation infrastructure investments (X) on a certain government's social objective function (W), we could seek to evaluate whether those investments were money well spent.

[1] Examples include Acemoglu (2010), Andrews, Gentzkow, and Shapiro (2020), Baum-Snow and Ferreira (2015), Einav and Finkelstein (2018), Finkelstein and Hendren (2020), Hansen and Heckman (1996), Heckman (2010), Holmes and Sieg (2015), Intriligator (1983), Keane (2010), Leamer (2012), Low and Meghir (2017), Manski (2013), Matzkin (1986), Nakamura and Steinsson (2018), Nevo and Whinston (2010), Paarsch and Hong (2006), Reiss and Wolak (2007), Rust (2014), Timmins and Schlenker (2009), and Wolpin (2013).

What can a researcher observe about this setting? As a starting point, we imagine that the policy variable X is observed for each member among a set of units of observation: for example, countries, regions, firms, or households. However, in general, the object of interest W is not observed—indeed, we are often interested in concepts, such as notions of economic welfare, whose measurement from even idealized datasets can be controversial.

It is at this point that the researcher's theoretical "model" enters the picture. Both the object of interest variable W and the policy variable X are related to other variables. First, the researcher believes that the object of interest W can be viewed as a function of an additional observable, an auxiliary outcome (or, at times, several outcomes) denoted by Y. We write this function as $W = g(Y,\theta)$. In this notation $g(\cdot)$ is a model—that is, it is a function the researcher will assume because there is reason to believe it is plausible—but the model's parameters, denoted by θ, may not be known. Second, we imagine that the researcher can observe an additional variable, denoted by Z, that is connected to X and is often referred to as an "instrument." This variable will be used to study the effect of X on Y in a manner discussed below but, as is already apparent, it is of no direct relation to the researcher's model or question of interest. Its role will be important, but merely instrumental—just as a hammer is an indispensable tool for hanging a painting on a wall, but the hammer itself is not much to look at.

To give a sense of how research from this framework might operate, one can imagine the researcher striving to assemble two ingredients. The first tells us how the researcher's policy of interest X affects the auxiliary outcome Y. The second tells us how the auxiliary observable Y translates into the unobserved outcome of interest W, a mapping that depends on the researcher's model $g(\cdot)$ and the parameters θ. This two-way breakdown is central to what follows.

While the discussion so far has been deliberately abstract, a number of essential points are already apparent. First, we are starting with a question—that is, how large is the change in W caused by a change in X?—and holding the question fixed. Second, since W is unobserved, we could not answer this question without the help of our model, whose role is to tell us how the observable Y relates to the desired outcome W. Third, since both Y and X are observable, it is possible, at least in principle, to use data alone to reveal the empirical effect of X on Y. Given knowledge of such effect sizes, the parameters θ are the only unknowns that stand in the way of the researcher arriving at an answer to the question posed.

Finally, and crucially, a researcher will typically have wide latitude to choose the set of Y variables being included in the model $g(\cdot)$. This is important because the logical essence of the model changes as we condition on more auxiliary outcomes—indeed, the strength of the assumptions being invoked in $g(\cdot)$ gets weaker as more outcomes Y are included. In this sense, the parameters of any model are specific to that model. As Fernandez-Villaverde (2008) puts it, in the context of procedures that use estimates of individual-level responses in aggregate-level models: "Borrowing parameters from microeconomic models forgets that parameters do not have a life of their own as some kind of platonic entity. Instead, parameters have meaning only

within the context of a particular model." While one could imagine economists building up a complete understanding of the world's economic parameters from the ground up—akin to the book full of natural constants that can be found in a chemistry lab—this isn't how most economics research actually works. We write down models that strike a balance between plausibility, parsimony, and (statistical) precision, but always relative to the question of interest and the data available. In this regard, it is no surprise to open an economics journal and find that almost any given empirical model will have a similar (and small) number of parameters to be estimated, regardless of whether they aim to reflect the Peruvian prawn industry or half of planetary production. The challenges of doing social science mean that the empirical metaphors that economists use are, unlike those in the field of chemistry, inevitably context-specific and deliberately parsimonious.

The Tyranny of Distance Between Data and Answers

We have set up the researcher's problem: a question to be answered, a set of available data, and a set of maintained assumptions that we call a "model." How can the researcher use these inputs of theory and data to answer the question that has been posed? I will build up one way of describing responses to this challenge in spatial economics, with examples along the way. These examples begin with settings in which the available data variation very closely answers the question of interest, and so the role of the researcher's model is relatively minimal. My examples then progress to settings with greater separation between data and answers, where the discussion will be organized around steps that researchers take to minimize such a distance—that is, to make the leap from data to answers under as plausible a set of theoretical assumptions as they can. These steps involve careful choices about which auxiliary outcomes Y to measure as well as an understanding of economic theory that helps inform the researcher's model.

As discussed above, the first key ingredient in all of the research described here will be the researcher's empirically grounded knowledge of how the policy of interest X affects some auxiliary outcome(s) Y. How can such knowledge be obtained? Thankfully, this problem is extremely well studied in the field of econometrics.[2] A key starting point is the researcher's belief that the instrumental variable Z satisfies an exogeneity restriction, one version of which amounts to the belief that the variation across units in this variable is as good as random. In some settings, this belief is easily justified. A good example of such a scenario would be when Z is literally a random variable, as with treatment assignment in a randomized controlled trial, where Z is a measure of whether a unit of observation in the trial received the trial's treatment or not. In other cases, the Z variable may draw on certain quasi-experimental features the researcher has isolated in a natural experiment. More

[2] See, for example, the textbook treatment in Angrist and Pischke (2009), and Matzkin (2013) and Chesher and Rosen (2020) for surveys of recent advances.

generally, the characteristics of the Z variable are such that the researcher is comfortable with an assumption of exogeneity.

When the instrument Z is exogenous, the researcher can faithfully "identify" (and hence, with a large sample, hope to arrive at an accurate estimate of) the magnitude of two causal relationships: how Z affects the policy variable X and how it affects the auxiliary outcome variable Y. Clearly, this information will be insufficient, in general, to answer the researcher's original question. There are two obvious problems: the desired outcome W is not (yet) a known function of Y, and Y is not (yet) a known function of X.

Nature's Bounty

Before continuing with the general case, we pause to discuss an idealized—though not uncommon—scenario. Suppose, first, that the parameter θ is known to the researcher. This amounts to saying that the outcome of interest W is a known transformation of the auxiliary outcome vector Y (often simply because $W = Y$, or because W is a known aggregation of individual-level values of Y). Second, the as-good-as-random instrument Z is similarly a known transformation of the policy variable X that features in the researcher's question (again, often because $X = Z$). Clearly, relative to the general case we started with, this researcher is in an extremely fortunate situation. But through painstaking effort and tons of ingenuity, some researchers have found themselves in exactly such a position, as the following example illustrates.

> *Example #1: Driven to Dhaka.* Rural laborers in low-income countries often face a choice between paying to migrate (perhaps seasonally) to a large city or working for an inferior wage on local farms. But how responsive are migration choices to changes in migration costs? Would a widespread reduction in travel costs cause sufficient migration that even local wages for those workers who stay behind might increase? Bryan, Chowdhury, and Mobarak (2014) randomly subsidized travel to a major city among a sample of rural households in Bangladesh in order to examine these questions. Akram, Chowdhury, and Mobarak (2017) followed up with a larger version that randomly varied the number of subsidized households per village. In both cases the policy of interest X (travel costs) was explicitly randomized (so in this setting, $X = Z$), and the outcomes of interest W (migration rates and village-level wages) were observed, so the effects of lower travel costs on migration and village-level wages were apparent. They were also surprising. Migration responses were enormous (and persisted even years after the one-off subsidy was gone) and wide roll-out within a village did raise local wages (despite the high density of nearby, untreated villages).

Nature's Instrumental Bounty

We now continue with a slightly less idealized setting. Continue to imagine that θ is known—so the researcher knows how to map from an observed auxiliary

outcome Y to the object of interest, W, perhaps simply because Y is in fact the object of interest. But now we retreat from the happy scenario in which the instrument Z is a transformed version of the policy of interest X. Effectively, the policy of interest X is no longer as good as randomly allocated in the researcher's dataset.

While challenging, this setting is familiar for economists. As discussed above, the effects of Z on Y and of Z on X are known, thanks to the exogeneity restriction. One more assumption—the so-called exclusion restriction—allows researchers to combine these two effects into knowledge of the desired effect, which is how X affects Y. The exclusion restriction requires that all of the effect of Z on Y happens because of the fact that Z affects X, rather than a potentially distinct (but ruled out, by assumption) effect of Z on Y even as X remains unchanged. It is conceptually distinct from the process that determines Z (which may underpin the researcher's belief in the exogeneity assumption), so it needs to be assessed on its own merits. Still, this assumption is plausible in many settings, and so the exogenous and excludable variation in instrumental variables plays an essential role in all of the work discussed below. The next example provides a setting in which it continues to be the case that $W = Y$, but no longer the case that X is the same as Z.

> *Example #2: Sealing the Suez.* How much would the GDP of a typical country be harmed if it were less open to trade? During the 1967–1975 Suez Canal blockade, caused by regional conflict, some shipping routes, such as Tokyo-Amsterdam, had to be redrawn while others, such as Tokyo–Los Angeles, were unperturbed. Feyrer (2021) argued that the resulting variation in the exposure of countries to the increase in shipping distances caused by the blockade could be used (as his instrumental variable Z) in order to estimate the effect of the blockade on the policy variable of trade flows (X) and on the outcome variable, GDP (in this case the auxiliary observable Y is the same as the outcome of interest W). Putting the two together implied that, for a typical country (among those affected by the blockade), when its level of trade openness (imports and exports as a share of GDP) fell by 10 percentage points, its real GDP per capita fell by about 5 percent.

Examples such as this and the previous one offer a compelling set of answers for the research questions posed, and these answers draw less on explicit theory than the work in the remainder of this article. Before going on, however, it bears stressing that it would be wrong to imagine that economic theory plays no role at all in studies like those discussed so far. To the contrary, researchers draw on theory when designing randomized trials, when justifying their belief in the exclusion restriction of an instrument Z being used, and even when making basic decisions such as which research questions to ask (which X's and W's to investigate among the infinite set of options) in the first place. In addition, there is often a desire to extrapolate beyond the lessons from any given study and thereby hope that the

estimates from one setting are "externally valid"—generalizable beyond the setting at hand—and theory provides an essential guide for doing so.

Surrogacy

Although researchers do sometimes find themselves in the fortunate position described in the previous two examples, most of the time they do not. The remainder of this article considers scenarios that feature such challenges. We continue to imagine that the researcher is using data and a valid instrument Z in order to establish the effect of the policy variable X on the auxiliary output variable(s) Y. But at this point, the researcher has gone as far as possible toward answering the basic research question without bringing in the model's theoretical assumptions. To bridge the gap between the observed auxiliary variable Y and the unobserved object of interest W, the researcher has no choice but to lean on the additional assumptions encoded in the function $W = g(Y, \theta)$. This abstract scenario exemplifies the inherent complementarity between theory and data. Without theory, the researcher would not be able to move from the auxiliary variable Y to the object of interest W. But without the quasi-experimental variation in the data, the theoretical assumptions needed would be far more ornate, time consuming, and subject to doubt if the researcher had to rely on theory, rather than data, for the empirical knowledge about the effect of X on Y.

To move from the observed Y to the unobserved object of interest W, even with the help of theory in the form of $W = g(Y, \theta)$, the researcher must know which values of the parameters in the vector θ to use in the mapping. We will now imagine that the researcher will draw on some additional data, labeled D, in order to arrive at an estimate of θ. Even though the details of this step are important, they vary across settings in ways that don't matter for the discussion here, so we shall summarize this estimation process as $D = \theta$. This implies that θ is known, thanks to the data elements embodied in D.

Summarizing the discussion so far, the researcher's model is a theoretical device for extrapolating (with the help of additional data, D) from the effect of X on Y, which is observed, to the effect of interest, of X on W, which is not observed. A simple illustration of this theory-as-extrapolation logic draws on what is referred to as a "surrogate" method in the field of statistics. Here is a classic example from the medical literature. A researcher is investigating whether a given cancer treatment drug (X in this context) improves patients' survival chances (W in this context). The researcher has enough experimental control to vary access to the drug X across members of the sample in an exogenous manner. However, measuring survival rates W is often impractical. For example, doing so may require waiting too long, or it may be the case that in-sample observations of survival rates W are just too noisy for researchers to hope to say anything conclusive about how the drug affects survival outcomes, given the sample size. Thus, the medical researcher's question cannot be answered without the help of theory and model-informed estimation.

In this case, the researcher's theory comes in the form of a model of physiology in which the mapping $W = g(Y, \theta)$, from various observable biomarkers or "surrogate

outcomes" Y to the survival rate W, is already well studied, to the point where $g(\cdot)$ and θ are known. Crucially, the biomarkers Y are chosen because they are far easier to observe than the survival outcomes W. So researchers use the "surrogacy assumption"—that is, the belief that their knowledge of $g(Y, \theta)$ and θ is correct—to use the inexpensive measurements of biomarkers Y to map from the policy variable X to the object of interest W. This research strategy effectively splits the job of empirical estimation into the two parts noted earlier: estimating the effect of the drug on the biomarkers (the effect of policy variable X on auxiliary outcome variable Y) and modeling the quantitative relationship between biomarkers and survival rates. In practice, this second part may simply involve estimating a linear relationship between the object of interest W and the auxiliary variables Y (in limited but vital settings where measurement of the object of interest W is feasible) but the principle generalizes to any potential mapping $g(Y, \theta)$.

While this may sound like an idealized setting found only in clinical medical trials, many studies in spatial economics share similar elements. The following is an example.

> *Example #3: Engel's Law meets Indian Trade Liberalization.* What effect did India's 1991 tariff liberalization have on the real income of households in regions that were specialized in sectors most affected by tariff reductions relative to households in regions that were not? Real income (W) is hard to measure in the absence of high-quality price data covering all consumption, a particular problem in this context, especially given the changes in product quality and variety that are emphasized as important mechanisms underpinning the gains from trade. To overcome this challenge, Atkin et al. (2020) describe primitive assumptions under which any cross-section of utility-maximizing households will obey an Engel's Law–like relationship: that is, as real household income increases, the share of income spent on, say, meat as a share of food expenditure declines in a stable and monotonic manner that is invariant to relative prices in non-food sectors of the economy (at least among those in which price measurement is difficult). These assumptions can then be invoked as a form of surrogacy assumption. The inverse of the estimated Engel-like curve relates the hard-to-measure desired outcome (real income, W) to the easy-to-measure surrogate (meat expenditure shares within food, Y)—and this estimated relationship thereby populates the parameters θ in $W = g(Y, \theta)$. Atkin et al. (2020) go on to exploit plausibly exogenous variation in the exposure of Indian regions to tariff reductions (Z), as previously developed by Topalova (2010). This method exploits an interaction between predetermined regional specialization across sectors and the Indian government's desire to homogenize variation in tariffs (as well as reduce the overall level), which meant that sectors with initially high tariffs had farther to fall in the 1991 liberalization cuts. Armed with Topalova's instrument, one can arrive at estimates of the effect of tariff exposure (X) on food budget shares (Y) and then use

the estimated Engel-based surrogacy relationship to estimate the effects on real income. In this way, Atkin et al. (2020) estimate large negative effects of the reduced import tariffs on rural households, evenly spread throughout both rich and poor rural households. In interpreting these results, the authors are careful to stress that relative effects across regions, not the overall effect on living standards in India as a whole, are the object of interest. The exogenous variation is cross-regional, so it cannot speak to the nationwide level effect.

Surrogacy-like assumptions often provoke skepticism in both the medical and economics literatures. But they often come with the ability for testing in special settings where W (and Y and X) are observed, because the implication of the surrogacy assumption is that X should have no effect on the difference between W and $g(Y, \theta)$.[3] In addition, the primitive economic assumptions that are invoked in the model $g(Y, \theta)$ may have additional predictions that can be tested.

More Challenging Extrapolation

In the classical surrogacy case, the researcher's model function $W = g(Y, \theta)$ is linear. Example #3 stressed a more involved case, but one that rested on the intuitive economic logic of Engel curves. In wider economics applications, the model function $g(\cdot)$ is often considerably more complicated. For example, the function $g(\cdot)$ could represent the solution to a large system of nonlinear equations that describes the general equilibrium of a competitive economy or the Nash equilibrium of a game-theoretic model of interactions between firms. It could even represent the result of a search over a set of feasible economic policies, where evaluating the merits of each candidate policy involves solving for the equilibrium that would be believed to prevail in an economy as a result of enacting the policy.

Whether $g(\cdot)$ is simple or complex, there is still substantial value in drawing a distinction between the two ingredients that the researcher will learn from the data: the effect of changes in the policy variable X on some auxiliary outcome Y, and the parameters θ that enter the model's mapping $g(\cdot)$. These two ingredients do not necessarily have the same provenance. By definition, θ does not connect neatly to some estimable effect of policy X in the researcher's own study—just like in the surrogates case, where θ must be drawn from a wider body of knowledge outside of the study at hand. The following example illustrates the power of extrapolation from estimated policy effects on auxiliary outcomes to a desired goal that involves feeding those estimated effect sizes into a more complex, equilibrium model.

Example #4: Trump's terms-of-trade war. How much would aggregate US real income change from levying import tariffs? How would matters differ if

[3]Athey et al. (2019) develop tools that weaken the assumptions behind surrogate methods, as well as those that allow a researcher to calculate bounds on the potential for bias due to, and to test for, violations of the surrogacy assumption.

foreign countries were to retaliate with their own tariff hikes? Fajgelbaum et al. (2020) study the tariff changes stemming from the 2018 trade war, instigated by the Trump administration, to answer these research questions. The researchers estimate effects of plausibly exogenous variation in US and foreign tariffs (so $X = Z$ here) on certain features of four key auxiliary outcomes (Y): prices and quantities of narrowly defined products coming into the US from tariff-hit countries relative to those that were spared; and similar prices and quantities for products leaving the United States for retaliating countries relative to others. These comparisons conveyed the striking finding that, despite the relatively large size of the United States in many global markets, tariff increases were immediately passed through into import prices, with large commensurate reductions in quantities crossing borders. While these results illustrate micro-level patterns of US and foreign supply and demand, an aggregate, general equilibrium model of entire US production and consumption $g(Y, \theta)$ is needed to answer the researchers' question about aggregate real income (W). To build such a model, Fajgelbaum et al. (2020) propose that US production is competitive and that production functions and inter-sectoral preference functions take the Cobb-Douglas form. Importantly, this model features producers who benefit from protective tariffs, producers who suffer from a rise in the price of imported materials, and consumers who both suffer from a rise in consumer prices and gain from an increase in tax revenue. Together, the model's assumptions imply both how θ can be pinned down by available data (D) as well as the mapping $g(Y, \theta)$ from the auxiliary outcomes Y to real income (W). Ultimately, the researchers' empirical model implies that the average US resident lost $22 of real income due to the tariffs (but these losses would have instead been gains, albeit very small ones of about $1, in a hypothetical scenario without foreign retaliation).

Sufficient Statistics

The discussion so far has emphasized the unavoidable need, when many questions of interest are concerned, to use theory embodied in the $W = g(Y, \theta)$ function to extrapolate from empirical knowledge of how the policy variable X affects the auxiliary variable Y to the question of interest—namely, how that same policy variable X affects the outcome W. Our image of theory as extrapolation raises the question: how "far" are we extrapolating?

One interpretation of "distance" relates to the "narrowness" of the space of reasonable economic assumptions under which $W = g(Y, \theta)$ is the correct—or equivalently, to the "width" of the space of reasonable assumptions under which this is the incorrect—model to use to answer the question at hand. Economists will have different perceptions of the magnitudes of these distances. Recall, however, that $g(\cdot)$ is not a conventional theoretical model, but an empirical model. That is, its content changes when values of the auxiliary variable Y are observed and when

the unknown parameters θ are pinned down by data. Thus, any assessment of the breadth of assumptions invoked by $g(\cdot)$ must be done while holding Y and the available data on other parameters θ fixed.

This distinction matters in practice. It is often the case that a researcher will consider using two different models that disagree on many things. However, the researcher may discover that the two models actually agree on what matters—that is, on their answers to the researcher's question of how changes in the policy variable X will affect the object of interest W—once we condition on features of the available data. Such features could derive from the estimated impact of the policy variable X on the auxiliary variable Y, and they could also derive from the values of the data that inform model parameters. Heckman (2010, p. 359) refers to this observation as "Marschak's maxim" in honor of Jacob Marschak (1953), who pioneered the understanding of situations in which low-dimensional combinations of model elements could suffice for answering a given policy question.

Another way of expressing this scenario is to say that, across the elements of some set of models, the evidence in the data acts as a "sufficient statistic" (or vector of statistics).[4] Conditioning on the available data is not just sufficient for filling in unknown elements of the model, in the usual sense of parameter estimation regarding the model's only unknown, θ. It may also be sufficient for eliminating elements of disagreement between two more plausible (but meaningfully distinct) models.

The endeavor to isolate sufficient statistics will depend on the question of interest. Asking models to agree on their answer to every question, even when we condition on available data, is a tall order. But asking models to agree when they are being used to answer a specific question is far more common and feasible. The following example illustrates the powerful logic of sufficient statistics in a spatial context.

> *Example #5: Million dollar or billion dollar plants?* When local governments offer subsidies and other incentives to attract large businesses, are their residents better off? Greenstone and Moretti (2003) describe a class of models in which workers are mobile and have identical preferences, local land is in fixed supply, other factors (such as capital) are mobile, and land markets are competitive. While the set of assumptions that defines this class is restrictive, it is far less restrictive than models that would go on to place additional restrictions on, or seek to estimate, the precise forms of firms' technologies (such as how those firms use mobile and immobile factors) and consumers' preferences (such as how consumers value the outputs made by firms and the public goods provided by local governments). Greenstone and Moretti (2003) then show that, within this class of models, paying a subsidy (X) to attract a business will impact local

[4] The application of sufficient statistics in this fashion has many parallels in other fields of economics. See Chetty (2009) and Kleven (2021) for reviews.

residents' welfare (W) by an amount that is equal to the observed change in land values in the location (Y)—that is, within this class, the auxiliary outcome Y is a sufficient statistic for W. Notably, this finding holds true despite the researchers' ignorance about the myriad complexities arising from general equilibrium product and factor market interactions (in this location and all others), local and wider externalities in production and amenities, and the gory details of how the subsidy is financed out of local funds (which may hence change tax rates and/or public service delivery) or supra-local sources. The intuition behind this result is that when one local factor (land, here) is in fixed supply and competitively exchanged, and yet all other factors are supplied perfectly elastically to a location, then the economic incidence of all location-specific phenomena (wages, prices, productivity, taxes, transfers, and others) would accrue to the fixed factor and be measurable via the observed change in its price. Based on this argument, and a plausibly exogenous source of variation in whether US locations narrowly win bids for a "million dollar" industrial plant (their instrument Z), the authors find that a typical winning location saw an increase in property values of at least $2.7 billion (in 2021 dollars), or about $11,000 per resident, within six years.

As compelling as this example is for answering the question of interest, it also serves to highlight the question-dependent nature of the sufficient statistics deployed. For example, it is harder to imagine how the estimates could be used to study the extent to which subsidies in one location are a zero-sum (or worse) game at regional or national levels, a topic of substantial policy interest (Slattery and Zidar 2020).

Necessary Statistics?

Once we identify settings where a class of plausible models agree—after conditioning on certain potential sufficient statistics—about the question at hand, the researcher has a stark choice to make. One option is to strengthen various theoretical assumptions so as to rule out models until only one model remains, and use that model alone. The alternative is for the researcher to find data on the sufficient statistic variables and make the model discrepancy go away. Such data will not always be available to researchers. But when it is, more and more research areas are transitioning to a view that the use of such data is no longer merely sufficient, but could also be considered necessary.

One example of this logic at work occurs in settings where the outcome of interest W corresponds to the value of the objective function of some decision-making agent who is believed to be optimizing (possibly subject to a constraint). This agent could be a consumer or a firm—and the next section discusses cases in which this agent may even correspond to the hypothetical representative agent of an entire economy. As economists know well (by the so-called envelope theorem of optimization theory), when an (optimizing) agent faces an exogenous change in

its environment, the first-order effect of that change on the value of its constrained optimization problem (W, here) is given by the direct effect of the change, because any indirect effects due to the agent changing its behavior are zero to first order. Crucially, this argument holds irrespective of the objective function. Thus, it can be applied even when the objective function that gives rise to W is not completely known, a natural scenario given our starting point that W is unobserved. The only knowledge required is the size of the direct effect of the change.

For the special case in which the change under consideration is to a set of prices faced by a consumer (known as Shephard's lemma), this result implies that the first-order proportional change in welfare is simply the product of any proportional price changes and the pre-change expenditure shares on the goods whose prices have changed. Thus, when the question of interest refers to a case in which the object of interest W is consumer welfare, a researcher can split up the analysis into two parts. First, the researcher could estimate the impact of the observed policy variable X on consumer prices Y. Second, the researcher could infer (to a first-order approximation) the effect of price changes on consumer welfare W with the help of data on all relevant initial expenditure shares. Formally, this approach would be invoking the assumption that (or choosing the model in which) the consumer under study is optimizing, and so as a result the effect of changes in consumer prices Y on consumer welfare W is fully revealed by the data on expenditure shares. This sufficient statistic result is useful because the space of reasonable models in which a consumer is just optimizing is extremely "wide" relative to the nested set of models in which the consumer is not just optimizing, but optimizing some particular utility function. The following example illustrates this idea at work.

> Example *#6: Pain and gain from tourists in Spain.* Who is helped and who is harmed when a location begins to export more? Allen et al. (2021) examine the recent doubling of tourist visits to Barcelona. They consider a class of models in which residents of any of the city's neighborhoods optimize a homothetic (but otherwise arbitrary) utility function over both their mix of consumption (including housing) goods from every neighborhood and their earnings from supplying labor to any neighborhood. Using data from Spain's largest consumer bank, the researchers observe data on individuals' budget and earnings shares for each of these options. These researchers therefore split up their analysis of how any individual's welfare (W) would be affected by a rise in (say) American tourists (X) into two components. First, they use plausibly exogenous variation in the timing and neighborhood concentration of certain tourists (Z) to estimate the effect of the change in tourism on prices and wages (Y) in each location. Second, they apply the insights above to argue that the effect of a given set of changes in wages and prices (Y) in any location on individual welfare (W) is a function of that individual's budget shares on each price and earnings shares on each type of income (D). These estimated effects imply that the tourism boom caused average welfare to rise for those in

peripheral city locations and to fall for those in the city center (where most tourism occurs).

Unnecessary Statistics

The discussion so far has imagined a researcher who wishes to answer an explicit counterfactual question using (because it is the only option) an explicit model. Further, the researcher has sought to leave as many of the details of that model as possible to be filled in by data features that can be conditioned upon.

One benefit of thinking this way is, as described above, the ability to minimize the extent to which the researcher's answers are driven by underlying assumptions. Another benefit is that the researcher may discover that the data requirements are actually simpler (and hence easier to collect) than may have first been apparent. Formally, this corresponds to a setting where the data requirement is a set of observable statistics that is actually a known combination of other observables. The most obvious version of this is where the data (on either Y or D) is a "macro-level" variable that is an aggregation over more micro-level statistics, as will arise when the micro-level statistics enter linearly and with uniform weight. This means that the long vector of micro data includes a set of *unnecessary statistics*, once we condition on observing the shorter vector of macro data. The following study describes an example where this logic applies.

Example #7: *Gains from trade in a gravity world.* How much does a country gain from the trading it does with the wider world? Arkolakis, Costinot, and Rodríguez-Clare (2012) consider a class of models in which consumers have constant-elasticity of substitution preferences, firms have heterogeneous but constant marginal costs of selling to any country, firms use one factor that is in fixed supply to each location, and firms compete either perfectly competitively or monopolistically competitively (with, in this latter case, fixed costs of developing a differentiated good and entering any market). In such an environment the welfare (W) cost of autarky (for example, by erecting prohibitive tariffs X) could range from zero to infinite depending on the heterogeneity in marginal and fixed costs. However, these researchers derive a surprising sufficient statistic result about a commonly used subset of models in this class known as "gravity" models—those that may differ in many underlying details but nevertheless display a constant and homogenous "trade elasticity," which is defined as the proportional change in a country's relative imports (which we could think of as an auxiliary outcome Y) from any two origins due to a proportional change in the relative tariff levied on those two origins (X). In particular, Arkolakis et al. (2012) show that the welfare cost of autarky for a given "Home" country is a function of just two statistics: the value of the trade elasticity and the current share of imports in Home's total consumption. These are both aggregate statistics, which implies that underlying

micro data, such as that on the sets of firms, products, and/or countries inside Home's aggregate import share, are unnecessary statistics for the question at hand and within the class of models considered. The same is true for the response of relative imports to relative tariffs—it is the aggregate value of imports Y that matters for learning the trade elasticity. As reported in Costinot and Rodríguez-Clare (2018), under these assumptions, for a country like the United States, the welfare cost of moving to autarky in 2011 is found to be 1.5 percent. This relatively low number arises both because the United States imports relatively little and the trade elasticity is thought to be relatively high.

Sufficient Functions

The language so far has stressed cases in which the sufficient statistic is either a single statistic or a vector of statistics. But nothing in the logic rules out cases where the sufficient statistic is actually an infinite-dimensional statistic—a sufficient *function*—that a researcher could hope to estimate (nonparametrically) in order to feed into the answer of the basic research question. At a high level of abstraction, this observation is trivial, because clearly the function $g(\cdot)$ is a sufficient function for answering the researcher's question. But a more common way for an economist to visualize the model is as a collection of functions—for example, the supply and demand systems for all firms and consumers in an economy. In this respect, the promise of a useful sufficient function is one that aggregates over (or otherwise combines) some or all of the many micro-level functions inside a researcher's model to arrive at the lowest-dimensional system that is needed to answer the researcher's question. Such a scenario implies not only the usual benefits of sufficient statistics—the ability to use data to avoid theoretical debate about the appropriate model to use within some wider class—but it can also serve as a guide to researchers about the minimal set of functions that are required to be learned from that data for the purposes of the goal at hand. The following is an example of such a case.

Example #8: *Gains from trade without gravity*. Let us return to the question from the previous example: How much does a country gain from trading with the wider world? Adao, Costinot, and Donaldson (2017) consider a class of models with arbitrary preferences and arbitrary technologies used under competitive conditions. Even though countries trade goods in these models (and in the real world), for every model in this class, and for the purposes of answering questions such as the one posed here, the model is isomorphic to one in which countries instead merely trade the services of their (geographically immobile) factors. Thus, any country has a set of well-behaved but "reduced" preferences over the factor services (rather than the goods) on offer around the world. Such reduced preferences for as-if factor service exchange, if known, can therefore summarize the underlying preferences and technologies for the goods in the world and

hence provide the inputs for welfare analysis. The underlying logic builds on that in Example #5: in general equilibrium, immobile factors are the objects onto which the total effects of other local phenomena accrue under competitive conditions. Putting this into practice, to the extent that there are fewer factors than goods, the summary offered by reduced preferences is dimension-reducing—and in the context of commonly used modeling environments with thousands (or even a continuum) of goods, the empirical dimension-reduction involved can be substantial. Adao, Costinot, and Donaldson (2017) use variation in transport costs (Z) to estimate reduced factor demand functions (relating factor service flows Y to trade cost shifters X). Based on such estimates, Costinot and Rodríguez-Clare (2018) calculate that the welfare (W) cost of autarky (a prohibitively high X) for the United States would be 2.3 percent (rather than the 1.5 percent mentioned in the previous example in the context of a gravity model).

Wedges, Welfare, and What-If Questions

The central theme of this article has been the interaction of economic theory and data. On the theory side, one of the most powerful ideas that economics has to offer is embodied in the first and second welfare theorems. These theorems state that, in the absence of market failures (such as externalities and market power), and with access to lump-sum transfers to address distributional concerns, along with some additional (and more technical) assumptions, the laissez-faire market allocation would be optimal.[5] The converse is equally important: in the presence of market failures, or in the absence of lump-sum taxation, market allocations are likely to be sub-optimal. This foundational theoretical result has important implications for the conduct of empirical work. Indeed, these implications resonate with many of the points raised above.

Designing Optimal Policies

Often, the researcher's counterfactual question will not just be "What would be the effect of a particular policy?" but "What is the policy that would be optimal in some well-defined sense?" How can researchers combine theory and data to answer questions such as these?

We shall begin by considering settings in which the object of interest W represents the welfare of an economy's representative agent—or equivalently, where the researcher believes it is plausible that policy could make (something close to) lump-sum transfers to agents as part of the optimal policy scheme. In such a setting, and in

[5] This statement assumes that all agents rationally pursue their best interest and so ignores policy motives deriving from a failure of agents to optimize. While such motives have featured in other branches of economics, they have seen far less focus in the areas I cover here.

the absence of market failures, the welfare theorems tell us that the optimal policy is already known: it is to step aside and let the market do its work. Obviously, in this case, there is no role for data or theoretical modeling to answer the question of interest. But the corollary is interesting: when the goal is to design optimal policies, the role that theory and data play is purely to provide a measure of the magnitude of market failures and of the consequences of real-world limits on lump-sum transfer schemes.

Consider, for example, the case of market failures. The intuition from the welfare theorems implies that optimal policy would align the prices that prevail in the actual economy with the "first-best" prices that would prevail in an economy that is equivalent—that is, an economy that features identical preferences, technologies, and endowments—but in which market failures are absent. Put differently, optimal policy would use taxes and subsidies to offset the wedges that market failures create between prices in the actual economy and those in the first-best equivalent. This framework provides the basis for proposals that call for imposing on polluters a tax equal to the wedge between the private and social cost of the pollutants they produce, or for offering subsidies to innovative producers equal to the wedge between the private and social benefits produced by their research and development efforts.

This implication of the welfare theorems is well known to economists. But it has a stark implication for the direction of empirical work on optimal policy of the sort described in this section: the goal of empirical work in such a context can focus on measuring the locus and magnitude of all relevant wedges and put other matters to the side.

How can such wedge measurement be done? We can generically think of market failures arising whenever the buyer of some "good" pays a different price for that good than the price that the seller receives. In some cases this is relatively easy to quantify, because the wedges are directly the result of taxes, subsidies, or other policies that leave a clear paper trail. For example, a 10 percent sales tax causes a clear distortion because whatever price the seller is charging for the good being exchanged, the consumer pays 10 percent more.

However, many of the wedges that concern spatial economists are not so easily observable. For example, consider the classic case of a factory that causes an externality when it expels pollution into a nearby river. Here, the "good" (technically, a "bad") changing hands is pollution, the "seller" of pollution is the factory, and the "buyer" of pollution is the nearby resident who drinks water from the river downstream of the polluting factory. Further, if the factory pays no penalty and bears none of the cost of its behavior, this pollution seller receives a price of zero when it sells this good. On the other hand, the buyer of the pollution is effectively (and involuntarily) paying a large price for the good because of the health damages caused by drinking polluted water. As before, the essence of this market failure is that the price the selling factory receives (zero) is different from the price that the buying residents are paying (large). But this wedge leaves no simple paper trail. Instead, it hinges on the (monetary equivalent of the) size of the health damages caused per unit of pollution. Nevertheless, the goal of wedge measurement in this case is clear: we need an estimate of the damage function relating health to pollution.

One way to measure wedges in these more challenging cases can be expressed as follows. Let one of the auxiliary outcome variables Y be an observed variable that measures the social benefit or cost of an agent's actions and let X denote an observed measure of the private benefit or cost, to that agent, of those same actions. As above, we imagine that the researcher has an instrument Z that allows estimation of the effect of X on Y. But such an effect is exactly a measure of the ratio of marginal social benefit to marginal social cost, which is the wedge we seek to understand. To take another example, consider the case of the markup (the ratio of price to marginal cost) that a firm with market power would charge. Here, the firm's action is the decision to produce more of its product. The marginal social value of this action, per unit produced, is simply the price it charges to consumers. The marginal private cost, to the firm, of producing is simply the cost of producing an additional unit. An estimate of the markup can be formed by estimating the treatment effect of the firm's production costs (at fixed input prices) on the firm's sales (at fixed output prices), as in Hall (1988).

Doing this for every wedge that seems relevant for the researcher's question is certainly challenging—even daunting. But a researcher can make substantial progress by replacing assumptions about wedges (including of course the assumption that they are all absent) with accurate measurement of wedges. The payoff of wedge estimation is particularly clear in the next example.

Example #9: Tennessee Valley Authority or Hudson Valley Authority? Where should place-based policies and infrastructure investments be optimally placed to maximize national output? Kline and Moretti (2014) evaluate the Depression-era investments (for example, in hydropower generation facilities) that were made in the Tennessee Valley. One clear benefit of such investments is that local firms and households had access to cheaper electricity, and perhaps the Tennessee Valley offered uniquely untapped engineering benefits as a place where new electricity generation capacity could be created relatively cheaply. But a more commonly voiced idea is that relatively underdeveloped areas, such as the TVA region, were places with untapped *economic* potential. Formally, this idea only makes sense if there are local positive externalities of production in the region—which would drive a wedge between private and social values of production and result in inefficiently low levels of output. Indeed, if such spillovers were higher in the Tennessee Valley than, say, in the Hudson Valley near Manhattan, then Tennessee would be a more efficient place to spend national investment funds than the Hudson Valley (at least on the margin). For this reason, Kline and Moretti (2014) devote significant effort to the estimation of the shape of the local spillovers (which will then govern the size of the wedge between social and private values of production at any location in the country); this amounts to estimating a non-linear relationship between local productivity (Y) and local size (X), using features of the TVA program as instruments (Z). Perhaps surprisingly, they find the local spillover

function to have approximately the same elasticity in all locations. This means that both small and large locations appear to have the same extent of externalities, and hence wedges, on the margin. It follows that (apart from engineering-related considerations) the Tennessee Valley Authority investments would have generated just as much additional national output wherever in the country they were targeted. The function relating national output (W) to the sizes of locations (X) appears quite flat—so when the question of interest concerns how best to use the TVA to manipulate X so as to maximize W, the answer is that almost any allocation would be equally good.

Continuing our theme of optimal policy design, a distinct motive for market interventions (beyond the attempts to offset market failures discussed above) may arise when the distributional goals underpinning our notion of policy optimality may not be feasible because lump-sum taxes and transfers are thought to be unrealistic. One alternative focal point in the public literature concerns the more plausible scenario in which a government can levy taxes in relation to a household's earnings only—in contrast to the case of lump-sum taxation in which any desired amount could be hypothetically taken from one household and transferred to another without affecting household decisions. Income taxation incurs inefficiencies because the government cannot condition tax liabilities on notions of effort (such as hours worked) or investment (such as time spent training) that households may make in the process of earning their income. In such settings, a government may wish to tax commodities (perhaps via import tariffs or location-specific business support) to achieve redistributional objectives, even in the absence of market failures.

An obvious challenge involved in incorporating such goals into empirical models of optimal policy design is that the analyst needs to know what the government's objectives actually are. For example, what weight should the government attach to the marginal consumption of a household below the poverty line, or to the top 1 percent of income earners? Economists are naturally disinclined to even dare to answer questions such as these. An alternative is to solve for the nature of optimal policies under any given set of conceivable weights, and offer a menu to the government to choose from, but this is usually impractical. However, the following example illustrates one way around this challenge.

Example #10: International trade and the equity-efficiency trade-off. How should import tariffs be designed to achieve redistributional objectives—such as to offset the distributional consequences that Autor, Dorn, and Hanson (2013) argue have resulted from the recent surge of US manufacturing sector imports from China? Costinot and Werning (forthcoming) work with a model in which the country of interest features no market failures and is small enough that it has no reason to impose tariffs in the hopes of improving its terms-of-trade (consistent with the evidence discussed in

Example #4). As such, the only motive for a tariff is that it may provide *pre*-distribution that cannot be achieved via income taxation. Costinot and Werning (forthcoming) also assume that the government's redistributional objective is a function of incomes (rather than other taxpayer identities) and that the observed income tax schedule reflects the government's redistributional objectives. In such a setting, these authors show that the optimal tariff on Chinese imports is a function of four sufficient statistics: the marginal income tax schedule, the income distribution, elasticities of labor supply at each income level, and estimates of the impact of Chinese imports on wages at each quantile of the income distribution. Remarkably, the optimal tariff, when written this way, does not depend on the government's preferences over the distribution of income, because these are already revealed by the observed tax schedule. Nor does it depend on the underlying economic details of exactly why Chinese imports might affect earnings differently across the distribution. To apply this formula, Costinot and Werning (forthcoming) use estimates of income quantile-specific wage (Y) impacts of Chinese imports (X) from Chetverikov, Larsen, and Palmer (2016)—researchers who themselves leveraged Autor, Dorn, and Hanson's (2013) empirical strategy of (analogously to the work in Example #3) comparing regions of America that had relatively greater employment in goods with greater Chinese import growth to regions with lower such exposure (as well as a measure of plausibly exogenous Chinese import growth derived from patterns of Chinese exports to other countries, Z). While the impact of Chinese imports is thought to have differed substantially across the income distribution, the implications of this finding for the redistribution-driven optimal tax on these imports is minimal, as the implied optimal tariff rate is less than a tenth of a percent.

Impacts of Other Shocks in the Presence of Market Failures

Finally, we consider now a researcher whose object of interest W corresponds, as above, to the welfare of a representative agent. But now the research question is not about optimal policy. Instead, we return to the case in which the researcher is studying the welfare effects of a change in some other characteristic X of the economic environment. For example, this X could be a change in the economy's technology, like the installation of new infrastructure.

What does the presence or absence of market failures imply for the researcher's answer to this question? As discussed above, when there are no market failures, and when lump-sum transfers are thought to be feasible, a consequence of the welfare theorems is that the market allocation is maximizing aggregate welfare. As a result, there are no first-order benefits from changing this allocation in response to an exogenous change in the environment. This observation implies—in a result known as Hulten's (1978) theorem, an economy-wide application of the envelope theorem that we mentioned earlier—that the first-order benefits of a shock to X in an efficient economy are given by the vector of "Domar weights" (which are the

value of production as a share of GDP) on all activities that are directly affected by X in the sense that the productivity-enhancing benefits of the shock occur in such activities. Furthermore, another implication is that even the second-order benefits are given simply by the *changes* in the Domar weights of directly-affected activities. Remarkably, the initial levels of, and endogenous changes in, prices and quantities of every other component of the economy do not need to be known to the researcher because they do not matter (up to second order).

These results may sound straightforward, but they have deep implications for empirical work in settings where researchers believe that distortions are limited. One is that an essential ingredient of any analysis will be the size of the direct productivity changes caused by the shock to X. This can be estimated in standard fashion: let Y represent the productivity of activities that X is plausibly directly affecting, and use the methods described earlier to estimate the effect of X on Y. Another is that, to the extent that first-order welfare changes suffice, the size of the Domar weights on those directly affected activities will be a set of sufficient statistics for the welfare impact. Finally, to additionally incorporate second-order welfare effects, it suffices to estimate the effects of X on an additional auxiliary outcome variable Y, namely the *changes* in the Domar weights. The following describes a classic example of this logic:

> *Example #11: Indispensable statistics and railroads.* How large are the economic benefits of massive investments in transportation infrastructure? Robert Fogel's (1964) landmark book, *Railroads and American Economic Growth: Essays in Econometric History*, examined the "axiom of indispensability"—that America's rapid growth in the late nineteenth century would not have happened without railroads. His analysis assumed that the economy was free of market failures and therefore focused on three goals. First, the reduction in the average user cost of transportation that the railroad network contributed relative to pre-existing transport system—this was Fogel's measure of the direct productivity benefits of railroads on the activity of transport, the directly affected activity. Second, the value of transported goods as a share of GDP before the railroad expansion—this was Fogel's measure of the Domar weight on transport. Third, the change in the value of transported goods over the time period in question—this was Fogel's measure of the change in the transport sector's Domar weight, as was necessary for quantifying the second-order welfare benefits. The methods that Fogel deployed did not apply econometric tools in the modern sense of the word. They focused on the total change in the amount of transport, and in the user cost of transport, over the period rather than an attempt to estimate the role of railroads in causing such changes. But Fogel's focus on these three indispensable statistics was clear, and it led to the provocative finding that the rail expansion increased GDP by no more than a few percentage points. As indispensable as the new technology of railroad may have looked to some observers, in an efficient economy railroads could

> not have been very beneficial unless they either drove large changes in user costs (which they probably did not), took place at a time when transport was a large sector in the economy (which it wasn't), or enacted substantial growth in the use of transport (which they probably didn't).

As powerful as Hulten's theorem can be, its logic breaks down in distorted environments. In writing about "Professor Fogel On and Off the Rails," David (1969) focused his criticism on the fact that Fogel's method was reliant on the controversial assumption that market failures were unimportant. In the presence of market failures, a first-order component of how changes in the feature X affect the object of interest W will now hinge on two additional mechanisms. The first concerns the extent to which the shock to X causes reallocations of primary factors toward those activities that have large positive wedges—that is, the activities for which social value exceeds private value. Such reallocations would generate a benefit of X that could not happen in an efficient economy. The second mechanism is the extent to which the shock actually changes the wedges themselves, which can provide additional benefits. Of course, some changes in X might have mixed or negative effects, perhaps mitigating certain market failures but exacerbating others.

This approach implies a feature of first-order welfare analysis that echoes Tolstoy's comment (at the beginning of *Anna Karenina*) that "[h]appy families are all alike; every unhappy family is unhappy in its own way." Here, all efficient economies are alike in their predicted responses of W to X (conditional on a given set of observed Domar weights). But every inefficient economy could see W respond to X in its own way (even after we condition on Domar weights). Predicting first-order welfare effects in undistorted economies hinges only on Domar weights. But doing so in distorted ones requires one to predict counterfactual reallocations, which requires strong modeling assumptions and measurement (of wedges and elasticities of agents' choice functions).

Baqaee and Farhi (2020) clarify and generalize these classical themes. Predicting the effects of counterfactual changes in X will require a full model of the economy of interest—at least any component of the economy in which reallocation could happen and in which wedges exist. For example, in a setting with only one factor of production—say, labor—and in which firms make goods that enter final output only, the extent of misallocation due to market failures in production is easy to see: it hinges on the extent to which some firms have larger value marginal products of labor than others. Further, if a shock of interest X were to have first-order reallocative effects on welfare W, it could do so if and only if it were to cause labor to move toward the firms with higher-value marginal products of labor; indeed, the first-order benefit of such a move is the size of the gap between value marginal products of labor (equal to the firms' relative wedges on output) times the change in labor reallocation that is due to the change in the policy variable X. At least in a simple setting like this one, modeling efforts would do well to focus on measuring pre-existing wedges—as per our previous discussion of optimal policy—and, just

as importantly, on understanding how the shock might be expected to cause labor reallocation across activities with different wedges.

Sometimes, the researcher's question concerns the welfare effects of shocks that have already occurred. In this case, the labor allocation is, in principle, an outcome we could observe—it is a *Y* variable—and we could use such observations to estimate the effect of our shock variable *X* on this particular *Y*. Then, the reallocative effects that underpin how the shock to *X* affects welfare (at least to first order) in this context would be given by simply the product of pre-shock wedges and our estimates of how changes in *X* affect *Y*. My final example pursues such an approach.

> *Example #12: Formalizing reallocation in Vietnam.* Can an export demand shock improve allocative efficiency? McCaig and Pavcnik (2018) study the large tariff reductions on US imports from Vietnam that followed from a 2011 bilateral trade agreement. Vietnamese manufacturing industries that saw relatively large reductions in US import tariffs exported more to the United States and expanded employment, and they did so relatively more among the formal-sector firms (as opposed to informal, household enterprises) within industries that could more feasibly overcome exporting hurdles. These reallocations would have no first-order welfare consequences if value marginal products of labor were equalized across and within industries, but one source of (within-industry) non-equalization derives from the fact that formal firms face greater taxation and regulation (and so would be expected to have larger value marginal products of labor). McCaig and Pavcnik (2018) estimate that such productivity difference wedges prior to 2011 were approximately 4 percent. They then quantify the effect of the trade agreement (*X*) on labor reallocation (*Y*) and multiply this estimated effect by the labor productivity wedge. The result suggests that aggregate labor productivity (*W*) rose by about 6 percent as a result of the trade shock.

Concluding Remarks

The article has offered an eclectic journey through some of the ways that recent work in spatial economics has sought to blend theory and data. Combining theory and empirics in this way is hard. Unsurprisingly, attempts to do so have been controversial. Skepticism stands in the way of those who wish to extrapolate from the estimated effects provided by quasi-experimental variation to the counterfactual questions that need to be answered. Yet given the necessity of such extrapolation, it seems vital that researchers understand the data-assumptions frontier in which they invoke only the most plausible theoretical assumptions necessary to map the data they have to the questions at hand, and in which they seek to minimize reliance on modeling assumptions by drawing on data that can resolve model ambiguities to the

greatest extent possible. The examples described above, drawn from a much wider body of work in the field, can be seen as pursuing that goal.

At the same time, spatial economists are witnessing a golden age of newly available sources of data. For example, troves of data tracking tax transactions, satellite imagery, smart phones, and credit card use are all being used to reveal spatial flows and linkages in previously unimaginable detail. The opportunities for blending the insights of economic theory with evidence from the spatial world around us have never been richer.

■ *I am grateful to Rodrigo Adao, Treb Allen, David Atkin, Arnaud Costinot, Gilles Duranton, and Enrico Moretti for helpful discussions about the themes of this article, and to the editors, Erik Hurst, Nina Pavcnik, Timothy Taylor, and Heidi Williams, as well as to Ben Faber, Pablo Fajgelbaum, and Amit Khandelwal, for their comments on an earlier draft.*

References

Acemoglu, Daron. 2010. "Theory, General Equilibrium, and Political Economy in Development Economics." *Journal of Economic Perspectives* 24 (3): 17–32.

Adao, Rodrigo, Arnaud Costinot, and Dave Donaldson. 2017. "Nonparametric Counterfactual Predictions in Neoclassical Models of International Trade." *American Economic Review* 107 (3): 633–89.

Akram, Agha Ali, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. 2017. "Effects of Emigration on Rural Labor Markets." NBER Working Paper 23929.

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

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro. 2020. "Transparency in Structural Research." *Journal of Business and Economic Statistics* 38 (4): 711–22.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arkolakis, Costas, Arnaud Costinot, and Andrés Rodríguez-Clare 2012. "New Trade Models, Same Old Gains?" *American Economic Review* 102 (1): 94–130.

Athey, Susan, Raj Chetty, Guido W. Imbens, and Hyunseung Kang. 2019. "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely." NBER Working Paper 26463.

Atkin, David, Benjamin Faber, Thibault Fally, and Marco Gonzalez-Navarro. 2020. "Measuring Welfare and Inequality with Incomplete Price Information." NBER Working Paper 26890.

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

Baqaee, David Rezza, and Emmanuel Farhi. 2020. "Productivity and Misallocation in General Equilibrium." *Quarterly Journal of Economics* 135 (1): 105–63.

Baum-Snow, Nathaniel, and Fernando Ferreira. 2015. "Causal Inference in Urban and Regional Economics." *Handbook of Regional and Urban Economics,* Vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 3–68. Amsterdam: Elsevier.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. 2014. "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh." *Econometrica* 82 (5): 1671–748.

Chesher, Andrew, and Adam M. Rosen. 2020. "Generalized Instrumental Variable Models, Methods, and Applications." *Handbook of Econometrics*, Vol. 7, edited by Steven Durlauf, Lars Hansen, James J. Heckman, and Rosa L. Matzkin, 1–110. Amsterdam: Elsevier.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics* 1: 451–88.

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

Costinot, Arnaud, and Andrés Rodríguez-Clare. 2018. "The US Gains from Trade: Valuation Using the Demand for Foreign Factor Services." *Journal of Economic Perspectives* 32 (2): 3–24.

Costinot, Arnaud, and Ivan Werning. Forthcoming. "Robots, Trade, and Luddism: A Sufficient Statistic Approach to Optimal Technology Regulation." *Review of Economic Studies*.

David, Paul A. 1969. "Transport Innovation and Economic Growth: Professor Fogel On and Off the Rails." *Economic History Review* 22 (3): 506–24.

Einav, Liran, and Amy Finkelstein. 2018. "Moral Hazard in Health Insurance: What We Know and How We Know It." *Journal of the European Economic Association* 16 (4): 957–82.

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

Fernandez-Villaverde, Jesus. 2008. "Horizons of Understanding: A Review of Ray Fair's *Estimating How the Macroeconomy Works*." *Journal of Economic Literature* 46 (3): 685–703.

Feyrer, James. 2021. "Distance, Trade, and Income — The 1967 to 1975 Closing of the Suez Canal as a Natural Experiment." *Journal of Development Economics* 153 (C): 1–12.

Finkelstein, Amy, and Nathan Hendren. 2020. "Welfare Analysis Meets Causal Inference." *Journal of Economic Perspectives* 34 (4): 146–67.

Fogel, Robert W. 1964. *Railroads and American Economic Growth: Essays in Econometric History*. Baltimore: Johns Hopkins University Press.

Greenstone, Michael, and Enrico Moretti. 2003. "Bidding for Industrial Plants: Does Winning a 'Million Dollar Plant' Increase Welfare?" NBER Working Paper 9844.

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

Hansen, Lars Peter, and James J. Heckman. 1996. "The Empirical Foundations of Calibration." *Journal of Economic Perspectives* 10 (1): 87–104.

Heckman, James J. 2010. "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature* 48 (2): 356–98.

Holmes, Thomas J., and Holger Sieg. 2015. "Structural Estimation in Urban Economics." In *Handbook of Regional and Urban Economics*, Vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange, 69–114. Amsterdam: Elsevier.

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

Intriligator, Michael D. 1983. "Economic and Econometric Models." In *Handbook of Econometrics*, Vol. 1, edited by Zvi Griliches and Michael D. Intriligator, 181–221. Amsterdam: Elsevier.

Keane, Michael P. 2010. "Structural vs. Atheoretic Approaches to Econometrics." *Journal of Econometrics* 156 (1): 3–20.

Kleven, Henrik J. 2021. "Sufficient Statistics Revisited." *Annual Review of Economics* 13: 515–38.

Kline, Patrick, and Enrico Moretti. 2014. "Local Economic Development, Agglomeration Economies, and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority." *Quarterly Journal of Economics* 129 (1): 275–331.

Leamer, Edward E. 2012. *The Craft of Economics: Lessons from the Heckscher-Ohlin Framework*. Cambridge, MA: MIT Press.

Low, Hamish, and Costas Meghir. 2017. "The Use of Structural Models in Econometrics." *Journal of Economic Perspectives* 31 (2): 33–58.

Manski, Charles F. 2013. *Public Policy in an Uncertain World: Analysis and Decisions*. Cambridge, MA: Harvard University Press.

Marschak, Jacob. 1953. "Economic Measurements for Policy and Prediction." In *Studies in Econometric Methods*, edited by William C. Hood and Tjalling C. Koopmans, 1–26. New York: Wiley.

Matzkin, Rosa L. 1986. "Restrictions of Economic Theory in Nonparametric Methods." *Handbook of Econometrics*, Vol. 4, edited by Robert F. Engle and Daniel L. McFadden, 2523–558. Amsterdam: Elsevier.

Matzkin, Rosa L. 2013. "Nonparametric Identification in Structural Economic Models." *Annual Review of Economics* (5): 457–86.

McCaig, Brian, and Nina Pavcnik. 2018. "Export Markets and Labor Allocation in a Low-Income Country." *American Economic Review* 108 (7): 1899–941.

Nakamura, Emi, and Jón Steinsson. 2018. "Identification in Macroeconomics." *Journal of Economic Perspectives* 32 (3): 59–86.

Nevo, Aviv, and Michael D. Whinston. 2010. "Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference." *Journal of Economic Perspectives* 24 (2): 69–82.

Paarsch, Harry J., and Han Hong. 2006. *An Introduction to the Structural Econometrics of Auction Data.* Cambridge, MA: The MIT Press.

Reiss, Peter C., and Frank A. Wolak. 2007. "Structural Econometric Modeling: Rationales and Examples from Industrial Organization." *Handbook of Econometrics,* Vol. 6, edited by James J. Heckman and Edward E. Leamer, 4277–415. Amsterdam: North-Holland.

Rust, John. 2014. "The Limits of Inference with Theory: A Review of Wolpin (2013)." *Journal of Economic Literature* 52 (3): 820–50.

Slattery, Cailin, and Owen Zidar. 2020. "Evaluating State and Local Business Incentives." *Journal of Economic Perspectives* 34 (2): 90–118.

Timmins, Christopher, and Wolfram Schlenker. 2009. "Reduced-Form Versus Structural Modeling in Environmental and Resource Economics." *Annual Review of Resource Economics* 1: 351–80.

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

Wolpin, Kenneth I. 2013. *The Limits of Inference without Theory.* Cambridge, MA: The MIT Press.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 211–222

Principles for Combining Descriptive and Model-Based Analysis in Applied Microeconomics Research

Neale Mahoney

At a fundamental level, there is no sharp distinction between descriptive and model-based empirical analysis. For example, while it is natural to think about the partial correlations estimated by ordinary least squares regression as descriptive, interpreting these estimates requires evoking, at least implicitly, a linear model of the underlying relationship. However, it is helpful to distinguish between different types of empirical analysis, and I find the descriptive versus model-based terminology useful. I will use the term "descriptive analysis" to describe empirical analysis when the primary goal is to summarize patterns in the data. I will use the term "model-based analysis" when the goal is to estimate an economic parameter, conduct a counterfactual, or make a statement about welfare.

In this article, I offer guidance on how to combine descriptive and model-based empirical analysis within a paper, drawing on my experience as a reader, author, and most recently a co-editor of applied microeconomics research. I will argue that it is important to construct a paper so that there is a tight link between the descriptive analysis and the bottom-line deliverable of the model-based analysis. To ground the discussion, I will begin with three recently published applied microeconomics papers: a health economics paper on prescription drug utilization, an education economics paper on school choice mechanisms, and a consumer finance paper on the pass-through of interest rates.

Drawing on examples from these papers, I will try to distill some lessons or principles. I will discuss the benefits of descriptive analysis, both for showing your

■ *Neale Mahoney is Professor of Economics, Stanford University, Stanford, California. He was a co-editor of American Economic Journal: Applied Economics.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.211.

identifying variation in a clear and intuitive way, and also for providing preliminary or partial evidence in support of your conclusions, even if your bottom-line conclusions require a quantitative model.[1] I will argue that you should clearly articulate the value-added of the model by explaining what you can learn from the model that cannot be learned from the descriptive analysis alone. I will also argue that you should use the descriptive analysis to guide your choices of what to model and what *not* to model. Finally, I will argue that you should choose parameters or counterfactuals that are informed by the identifying variation in the data and use descriptive analysis to help the reader form a prior belief over the parameter estimates and counterfactuals that follow.

For most of this essay, I will assume that you have decided to write a paper that starts with descriptive analysis and then proceeds to model-based analysis. This is not the only—or necessarily the best—way to craft a paper. Toward the end of the essay, I will share some thoughts on when this ordering may be desirable. I will also offer a perspective on viewing research in applied microeconomics as offering a set of trade-offs, in which the researcher needs to justify additional model-based assumptions in terms of the additional insights they deliver.

Three Running Examples

I will work with three examples, drawn from papers that span three fields within applied microeconomics: health economics, education, and consumer finance. These papers take very different approaches to the descriptive and model-based analyses. I have also chosen papers written by people I know—including a paper where I was a co-author. The reason is that I wanted to have frank conversations with the authors about the reasons behind the choices they made. Here I provide brief summaries of the papers, focusing on the connections between the descriptive and model-based analyses that are the focus of this article.

The "Donut Hole" in Medicare Part D

The Medicare Part D program provides insurance for drug expenditures for the elderly in the United States. However, for many years the insurance contract had an infamous "donut hole": consumers were subsidized by the program up to a lower level of annual expenditures and above a higher level of annual expenditures—but there was a region in the middle of the contract (the donut hole) where consumers had to pay the full cost of drugs out of pocket. Einav, Finkelstein, and Schrimpf (2015) study how spending on prescription drugs is influenced by the nonlinearity in incentives created by the donut hole. They also use the variation from this nonlinearity to study consumer behavior more broadly and to estimate the impact of counterfactual insurance contracts.

[1] I have intentionally avoided the reduced form versus structural terminology. These words have a precise meaning in certain contexts, and I do not want to risk confusion by using them imprecisely.

In the first part of the paper, the authors use the sharp jump in the out-of-pocket price at the "kink" in the contract at the start of the donut hole to generate visually compelling descriptive evidence. They show substantial "bunching" of annual spending at the kink, which allows them to reject the null of no response to incentives. They also show that the probability of a new drug purchase decreases as spending reaches the kink, with stronger impacts in December than earlier months in the year. The anticipatory response prior to December shows that people are forward-looking, while the stronger effects in December indicate that either uncertainty or partial myopia limits the responses in earlier months.

In the second part of the paper, the authors build a dynamic model of drug utilization, which allows for a stochastic health process, price sensitivity, and (partial) forward-looking behavior. The estimated model allows the authors to go beyond the qualitative evidence on bunching and quantify the response to the nonlinear incentives created by the donut hole by comparing outcomes under the observed nonlinear contract to the ones resulting from a linear counterfactual contract. The model also allows the authors to quantify behavior in terms of economic parameters, such as a weight the consumers place on future outcomes in their decision-making. Appealingly, the model is estimated via generalized method of moments to match the bunching patterns documented in the descriptive analysis.

High School Choice in New Haven

The New Haven, Connecticut, school district has offered students a mechanism for choosing between high schools since the 1990s. Such school choice mechanisms are used in many cities, and they raise some common concerns. Are students better off listing their actual choices, even knowing that certain popular schools will be oversubscribed with those listing it as a first choice? Or should students instead play a strategic game in which they put a second or third choice at the top of their list, in the belief that they will have a better chance of actually getting into that school if they list it first? At the time of the study, New Haven had implemented a mechanism that, by allowing applicants to express the intensity of their preferences, rewarded strategic play. The tradeoff is that if students are informed and sophisticated, the ability to express intensity of preferences in the New Haven school choice mechanism can improve welfare relative to a "strategy-proof" mechanism which only rewards listing one's true rank-order preferences. However, if students are uninformed or unsophisticated, the New Haven mechanism could lead to lower average and less equitable outcomes.

Kapor, Neilson, and Zimmerman (2020) study how the accuracy of beliefs affects the welfare from different school choice mechanisms. In the first part of the paper, the authors describe results from a survey of the school preferences of 417 students, combined with data on how students listed their school choices on their administrative application forms. They document that 32 percent of students are "revealed strategic," in the sense that they did not list their most preferred school first in their submission to the school district. However, the authors also show that this strategic behavior is poorly informed. A descriptive analysis shows that half of the revealed strategic students are "mistakenly strategic," in the sense that they

would have been better off listing their preferred school first rather than strategically listing another school first. Based on survey responses, students often hold beliefs that differ substantially from rational expectations about the probabilities of admission: for example, they are on average highly optimistic about their admission probabilities at second-ranked schools.

In the second part of the paper, the authors build a model of high school admission applications that allows for beliefs to diverge from rational expectations as documented in the descriptive analysis. A key decision is how to model subjective beliefs. The descriptive analysis provides no evidence of strategic information acquisition, so the authors do not allow subjective beliefs to vary with preferences. Instead, motivated by the descriptive analysis, the authors allow the wedge between subjective and rational beliefs to vary with the rank of the school chosen, priority of schools, and idiosyncratic school and individual components.

The authors use their estimated model to conduct counterfactuals that connect directly to the results from the survey. For example, the authors show that in a situation where subjective and rational beliefs diverge, a (counterfactual) strategy-proof mechanism would achieve higher welfare and improve equity. Indeed, the authors show that one needs to eliminate nearly all of the wedge between subjective beliefs and rational expectations for the New Haven mechanism, with its additional ability to express intensity of preferences, to be preferable on welfare and equity grounds. Finally, the authors show that if researchers didn't account for subjective beliefs and assumed rational expectations when estimating their model, they would have erroneously concluded that the New Haven mechanism was superior.

Pass-Through of Lower Interest Rates for Banks into Increased Borrowing by Consumers

Central banks, such as the Federal Reserve, can stimulate the economy by providing banks with lower-cost capital and liquidity. The idea is that these lower costs will encourage banks to expand credit to consumers who will, in turn, increase their borrowing and spending. Agarwal et al. (2017) argue that the impact of a reduction in banks' cost of funds on aggregate borrowing can be decomposed into the product of banks' marginal propensity to lend to borrowers and those borrowers' marginal propensity to borrow, aggregated over all borrowers in the economy. They study how frictions, such as asymmetric information, affect the pass-through of lower interest rates for banks into increased borrowing and spending by consumers. They apply this framework by estimating heterogeneous marginal propensities to borrow by consumers and marginal propensities to lend by banks in the US credit card market.

In the first part of the paper, the authors directly estimate consumers' marginal propensity to borrow using quasi-experimental variation in credit limits. Banks sometimes set credit limits as discontinuous functions of consumers' credit scores. For example, a bank might grant a $2,000 credit limit to consumers with a credit score below 720 and a $5,000 credit limit to consumers with a credit score of 720 or above. The authors identify 743 credit limit discontinuities in their data, located

across the credit score distribution, and use these discontinuities to estimate heterogeneous marginal propensities to borrow for consumers with different credit scores.

In the second part of the paper, the authors turn to estimating the marginal propensity of banks to lend to different customer groups. Estimating the marginal propensity to lend in a direct way using observed changes in banks' borrowing costs is challenging, because such changes are typically correlated with shifts in the economic environment that also affect borrowing and lending decisions. The authors write down a model of optimal credit limits to show that a bank's marginal propensity to lend depends on a small number of sufficient statistics that capture the relationship between changes in lending and profits. These sufficient statistics can be estimated using the same credit limit discontinuities, allowing the authors to recover heterogeneous marginal propensities to lend to borrowers with different credit scores. The authors show that bank lending is close to the optimal level implied by the model, providing support for the modeling assumptions.

In the final part of the paper, the authors combine the model-free estimates of consumers' marginal propensity to borrow with the model-based estimates of banks' marginal propensity to lend. They then use these estimates to describe the strength of this bank lending channel and show how features of the economic environment, which influence the marginal propensity to borrow and to lend, affect the strength of this channel.

Five Principles

In this section, I discuss five principles for combining descriptive and model-based analysis, as illustrated by the three papers summarized above.

1. Show Your Variation with Descriptive Analysis.

Many applied microeconomics papers are built around an empirical approach (sometimes referred to as a research design). For this type of paper, a primary goal of the descriptive analysis is to "make the case" for the identifying variation that drives the rest of the empirical analysis. Broadly, your aim should be to explain where your variation comes from, show that it is powerful, and show that it is valid.

The right way to show the variation depends on the context. In the Medicare Part D paper, the key source of variation is the donut hole that exposes beneficiaries to increased out-of-pocket costs. The authors show that the donut hole can be characterized by a kink in the contract that maps drug spending to out-of-pocket costs. They describe and visually illustrate the donut hole in the standard insurance contract, and in the non-standard contracts that they also use in their analysis.

In the credit card paper, the key variation is the jump in credit limits at specific credit scores, which the authors take advantage of by using a regression discontinuity design. To explain and illustrate this variation, the authors provide institutional context on how bank underwriting models give rise to these types of jumps in credit limits and provide visual examples of the discontinuities in their data. They then

establish the validity of these credit limit quasi-experiments by showing that other factors trend smoothly through the discontinuities and show there is no evidence of bunching above the discontinuities.

2. Use the Descriptive Analysis to Provide Preliminary Evidence.

As an author of applied microeconomics research, you should also use the descriptive analysis to provide preliminary or partial evidence for the paper's conclusions, while recognizing that the bottom-line conclusions will require a quantitative model.

For instance, in the Medicare Part D paper, the authors show visually compelling evidence of bunching around the kink (and show that the location of this bunching moves as the kink moves across years). This evidence allows the authors to reject the null hypothesis that there is no response to incentives, but the descriptive evidence is only partial in the sense that it does not allow the authors to quantify whether the response should be considered "large" or "small" in magnitude.

In the school choice paper, the authors provide evidence that students are "revealed strategic" in how they list schools, but are simultaneously "mistakenly strategic" in the sense that they would sometimes have been better off if they had listed schools in a non-strategic way. This indicates that mistakes may be important, but without further modeling assumptions, it cannot fully establish the quantitative importance of these mistakes.

Choosing how much and exactly what descriptive evidence to show is a balancing act. Weak or irrelevant descriptive evidence is a waste of time and can create problematic first impressions. At the same time, some readers may find the basic descriptive evidence more credible than model-based results, and you do not want to shortchange these readers. Getting feedback in seminars and conferences is useful for striking the appropriate balance.

3. Use the Descriptive Analysis to Guide Choices of What to Model—and Not Model.

Another key function of the descriptive analysis is to guide and support modeling choices. In the Medicare Part D paper, the authors show that consumers respond to the donut hole before the end of the year, but to a lesser extent than their response at year's end. These facts motivate the specification of a model where consumers are forward-looking, but potentially not fully so. In the school choice paper, the authors present descriptive evidence that suggests that mistakes are the result of poor information. Based on survey responses, students often hold beliefs that differ substantially from rational expectations admission probabilities: one example, as noted, is that they are on average highly optimistic about their admission probabilities at second-ranked schools. This motivates the decision to model mistakes as arising from mistaken beliefs, as opposed to another mechanism.

In my view, a signal benefit of a paper that starts with descriptive analysis and then presents the model is that you can use the descriptive evidence to justify what *not* to model. In this way, your modeling choices can be more transparent and less

arbitrary, without sacrificing the ability to capture key features of the environment. For instance, in the school choice paper, a natural consideration is whether people engage in strategic information acquisition—that is, whether they acquire more or better information about schools in their consideration set. In the descriptive analysis, the authors do not find that students have better information about the schools they are considering. Rather than falling into the trap of extending the model because of convention or because an extension would be "cool," the descriptive analysis provides the authors with evidence to justify their decision *not* to model strategic information acquisition—so that they can focus on what matters in their setting.

4. Clearly Articulate the Value-Added of the Model.

As mentioned at the start, I believe it is useful to think about the model as offering the reader a trade-off: If the reader is willing to accept the assumptions embedded in the model, then you can deliver additional and more economically relevant results.

In the Medicare Part D paper, the authors use the limitations of the descriptive analysis to motivate the model. In particular, they describe how the descriptive evidence on bunching allows them to *qualitatively* establish that there is a response to incentives but does not allow them to *quantify* the magnitude of this response. To gauge the economic magnitude of the response, and to gain a deeper understanding of partially forward-looking consumer behavior, they need to know how people would have behaved under a counterfactual linear contract without a donut hole. Because of the dynamic nature of behavior, estimating such counterfactual behavior requires a model.

In the credit card paper, the authors can recover consumers' marginal propensities to borrow in a model-free way using the credit limit discontinuities. However, recovering banks' marginal propensities to lend from time series data is difficult because shifts in banks' cost of funds—which are the result of policy actions by the central bank—often occur precisely when the economic environment is rapidly changing.[2] This motivates their model-based approach, in which they use a small number of sufficient statistics to pin down the lending propensities. They argue that the assumptions underlying this model-based approach—that bank lending responds optimally to changes in the cost of funds and that they can measure the incentives faced by banks—are reasonable in their setting.

The bottom line is that you need to engage in an act of persuasion and sell the model to the reader. To do so, you want to clearly articulate that the value-added of the model is high, in that it delivers considerably more insight than the descriptive analysis alone.

[2]For example, there was a large drop in the cost of funds for US banks in fall of 2008, when in response to the financial crisis the policy interest rate of the Federal Reserve (the federal funds rate) was set to near-zero. However, this was exactly the period when lenders and borrowers were updating their expectations about the economy, making it hard to separate out the effects of the drop in the cost of funds.

5. Choose Parameters of Interest and Counterfactuals That Are Informed by Your Variation.

Having specified and estimated a model, the final part of many papers discusses parameter estimates or conducts counterfactuals. The goal here is to deliver analysis that is more economically relevant than what could have been learned from the descriptive analysis alone—but is still informed by the data. Both of these are important. To get the reader to accept stronger assumptions, you need to be able to offer more economically relevant outcomes. At the same time, the results will be more credible if there is a tight link between the underlying variation presented in the descriptive work and the parameters or counterfactuals delivered by the model.

For instance, in the Medicare Part D paper, the main counterfactual is the effect of removing the kink. This comparison is clearly economically relevant: it is the natural benchmark to gauge the effect of the kink and it was a frequently discussed—and eventually implemented—policy reform. Since the descriptive analysis shows bunching, it is closely connected to the variation in the data.

In the school choice paper, the focus of the model is to incorporate inaccurate beliefs—and the resulting mistakes—into a state-of-the art school choice model. With model-based estimates of inaccurate beliefs in hand, the authors can then examine the effect of a counterfactual strategy-proof mechanism—and examine the effects of correcting beliefs holding the choice mechanism fixed. The counterfactual mechanism with correct beliefs helps quantify the cost of mistaken beliefs that is identified in the descriptive analysis, while the strategy-proof mechanism shows the benefits of a practical solution to the problem of inaccurate information. Indeed, the New Haven schools have now, with the researchers' help, rolled out a version of a strategy-proof mechanism.

The credit card paper uses the model and evidence from the quasi-experiments to recover banks' marginal propensities to lend. The marginal propensities to lend, combined with the directly estimated marginal propensities to borrow, allow the researchers to recover the pass-through of changes to banks' cost of funds. The heterogeneous estimates of banks' marginal propensities to lend are closely connected to the prior descriptive analysis, using the same quasi-experiments that are used in the model-free analysis to estimate consumers' marginal propensities to borrow.

More generally, it's important to emphasize that counterfactuals or discussion of economic parameters shouldn't be an afterthought, completed at the eleventh hour before a presentation or submission deadline. Choosing counterfactuals that provide economically relevant insights that go beyond what you could learn from the descriptive analysis but are still informed by your data—that use but don't abuse your model—requires careful thought and consideration. Don't sell yourself short.

Data-Then-Model or Model-Then-Data?

For most of this essay, I've taken as given that an applied microeconomics research paper should start with descriptive analysis and then proceed to model-based

analysis. However, an obvious meta-question is whether a data-then-model or model-then-data ordering is preferable.

Choosing how to structure a paper can be difficult—and I don't think there is always a right choice. Editors and authors sometimes disagree about the appropriate ordering, and my coauthors and I have sometimes switched the ordering during the course of a project. There are also more complex organizational structures—such as the use of an illustrative toy model, descriptive analysis, and then a richer econometric model—that I will not delve into here. With these caveats in mind, here are some thoughts that can help inform this decision.

It can be preferable to lead with a model when you need a model to guide decisions on what data to collect. Consider a field experiment where you collect your own survey data. For such a project, you would ideally use model-based reasoning to guide your decisions on what questions to ask in your survey. When writing the paper, it may be useful to present the model first to help motivate and justify the survey design.

Similarly, it can make sense to start with the model when the data is non-standard and you need the model to provide guidance on what sort of basic data analysis to conduct. For instance, if you have social network data, it may be hard to summarize the structure of the social network before introducing a model that can help define measures of network structure.

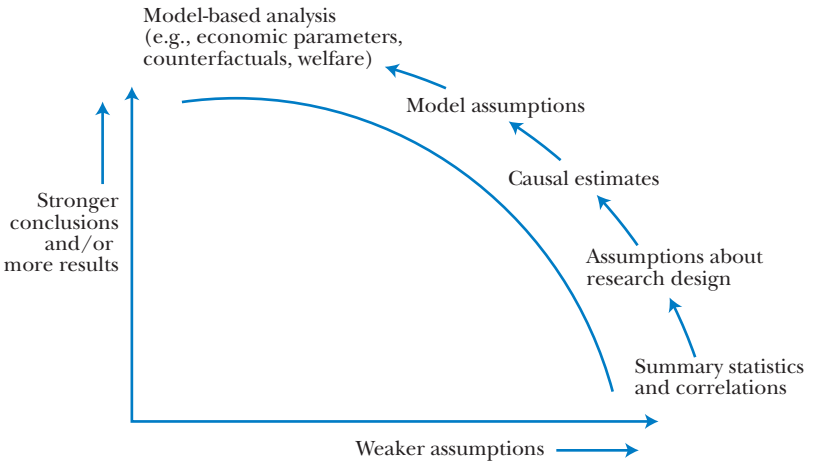
It can also be advisable to start with the model when the conceptual idea imbedded in the model is the main contribution of the paper. For instance, if your paper is proposing a new economic mechanism, then it is natural first to present the model that lays out this mechanism, and then present the data analysis that allows you to quantify its importance.

Conversely, one reason that it can be useful to lead with the data analysis arises if you want to use facts in the data to guide the modeling choices. For instance, in the Medicare Part D paper, the decisions of what to emphasize in the model—price sensitivity, uncertainty, forward-looking but not perfectly forward-looking behavior—are motivated by facts uncovered in the descriptive analysis. Similarly, in the school choice paper, the decision to write down a model with inaccurate beliefs—along with the specific decisions on how to model the wedge between subjective beliefs and rational expectations—would have been hard to motivate without the preceding descriptive analysis.

A second reason to use a data-then-model structure is that it keeps more readers engaged with your paper for longer. For better or for worse, I suspect that readers of applied microeconomics research are more likely to be "turned off" by a model than by descriptive analysis. If you lead with your model, you may lose some readers fairly early in the paper; whereas if you start with the descriptive analysis, you're more likely to retain your readers for at least some of your findings, even if you still lose them when you get to the model section.

A third appeal of the data-then-model ordering is that it is often a better reflection of the research process. Based on experience and conversations with colleagues, my sense is that many applied microeconomics researchers conduct

Figure 1
The Frontier Between Strength of the Assumptions and More Economically Relevant Results



Note: Figure depicts the trade-off between the strength of the assumptions and more economically relevant results.

extensive descriptive analysis before undertaking the effort of specifying and estimating a structural model. While papers should not be written as a chronology of the research process, ordering the paper in the same way in which the research was done often comes across as more natural.

Taking a step back, a metaphor I find useful is the exploration of a decision tree. In constructing a paper, it is smart to lead with the analysis that most quickly and efficiently prunes branches from this tree. If there is an overwhelming number of possible branches of data analysis, it may be more natural to start with the model to guide which branches to explore. If there is a rich set of models that could be plausible, it may be more useful to start with the data analysis to narrow the scope of the modeling exercise.

Concluding Thoughts

It is useful to think about data-then-model papers as tracing out a frontier that trades off the strength of the assumptions for more economically relevant results, as shown in Figure 1. At each stage in the paper, you are offering the reader a deal: if you accept some additional assumptions, then I will provide you with additional results. If the reader is willing to accept assumptions about the validity of the empirical approach, you can offer causal estimates. If the reader is

willing to accept additional assumptions about the economic environment, you can deliver additional results in terms of economic parameters, counterfactuals, or welfare.

Economist-readers understand trade-offs, and my sense is that they will be more likely to accept model-based assumptions if the paper is structured in a way such that they know they are getting something in return. In addition, economists have highly heterogeneous preferences about the kinds of model-based assumptions with which they are comfortable. This type of structure allows the reader to situate themselves at the point on this frontier that best matches their preferences—and allows the reader to "get off the train" at the point where they are no longer comfortable with the trade-off being offered.

■ *This paper was completed before I took leave to work at the White House National Economic Council. I thank my frequent collaborators Liran Einav and Amy Finkelstein for numerous conversations that have shaped my thinking on this topic. They deserve no blame for any faults that remain.*

References

Agarwal, Sumit, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel. 2018. "Do Banks Pass through Credit Expansions to Consumers Who Want to Borrow?" *Quarterly Journal of Economics* 133 (1): 129–90.

Einav, Liran, Amy Finkelstein, and Paul Schrimpf. 2015. "The Response of Drug Expenditure to Nonlinear Contract Design: Evidence from Medicare Part D." *Quarterly Journal of Economics* 130 (2): 841–899.

Kapor, Adam J., Christopher A. Neilson, and Seth D. Zimmerman. 2020. "Heterogeneous Beliefs and School Choice Mechanisms." *American Economic Review* 110 (5): 1274–315.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 223–244

Overreaction and Diagnostic Expectations in Macroeconomics

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer

Dynamic macroeconomics is one of the great accomplishments of twentieth century social science. It recognizes the centrality of forward-looking behavior for investment, consumption, and other major decisions of consumers and firms. The bedrock assumption of this research program is that expectations are "rational," meaning that decision-makers make optimal use of available information when making their forecasts. Indeed, this research program is often referred to as the "rational expectations revolution" (Lucas and Sargent 1981).

Despite the success of dynamic macroeconomics, growing evidence using surveys rejects any pure version of the rational expectations hypothesis (Souleles 2004; Vissing-Jorgensen 2003; Mankiw, Reis, and Wolfers 2003). To account for some of this as well as other evidence, early models maintained rational belief formation, but introduced costs of acquiring or processing information (Sims 2003; Woodford 2003). This approach has proved useful to explain sluggish price movements (Mankiw and Reis 2002). Recent evidence, however, points to deeper departures from rationality, which violate basic laws of conditional probability. The expectations of professional forecasters, corporate managers, consumers, and investors appear to be systematically biased in the direction of overreaction to news (Bordalo

■ *Pedro Bordalo is Professor of Financial Economics, Saïd Business School, University of Oxford, Oxford, United Kingdom. Nicola Gennaioli is Professor of Finance, Bocconi University, Milano, Italy. Andrei Shleifer is Professor of Economics, Harvard University, Cambridge, Massachusetts. Their email addresses are pedro.bordalo@sbs.ox.ac.uk, nicola.gennaioli@unibocconi.it, and ashleifer@harvard.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.223.

et al. 2020). As a result, beliefs are too optimistic in good times and too pessimistic in bad times, at the individual level and sometimes at the consensus level as well.

In this paper, we present the case for the centrality of overreaction in expectations for addressing important challenges in finance and macroeconomics. We begin with a brief overview of several formulations of expectations considered by economists. We then make three arguments. First, non-rational expectations by market participants can be measured and modeled in ways that address some of the key challenges posed by the rational expectations revolution, most importantly the idea that economic agents are forward-looking and form beliefs using their models of the economy (Muth 1961; Lucas 1976). We, among others, have constructed models of forward-looking but overreacting expectations, such as "diagnostic expectations" (Bordalo, Gennaioli, and Shleifer 2018). These models can be estimated using survey data and integrated into dynamic macroeconomic analyses.

Second, belief overreaction can account for many long-standing empirical puzzles in macro and finance, which emphasize the extreme volatility and boom-bust dynamics of key time series, such as stock prices, credit, and investment, in a natural and empirically tractable way. In essence, excess volatility and predictable boom-bust cycles arise because expectations overreact to news and are subsequently systematically corrected. The mechanism of overreaction in beliefs links excess volatility of stocks to return predictability, credit market frothiness to increased risk of financial crises, and macro financial booms to subsequent recessions.

Third, overreaction has two important advantages over conventional mechanisms used in economic models to produce excess volatility: it relies on psychology and is disciplined by survey data on expectations. We briefly discuss frequently used mechanisms that seek to maintain rational expectations, including exotic preferences and long-run risk. We assess the predictions of these models critically in light of the available survey evidence. Relaxing the assumption of rational expectations seems like a better research strategy, both theoretically and empirically.

A Very Brief History of Expectations Research

Before the rational expectations revolution, survey expectations were a central part of standard macroeconomic analysis. Starting in the 1940s, the National Bureau of Economic Research published several volumes on data on market participant forecasts, such as *The Quality and Significance of Anticipations Data* (1940). Although these early studies presented no systematic analysis of the structure of forecast errors, they were informed by a model of beliefs called adaptive expectations (Cagan 1956). This formulation was backward-looking, with expectations modeled as a distributed lag of past changes, with fixed exogenous coefficients. This formalization yielded initial sluggishness of beliefs. After a long period of price stability in goods or financial markets, expectations of future prices would remain anchored, despite growing prices, so that beliefs would only slowly adjust to the new regime. In the presence of positive feedback mechanisms, such as wage renegotiations feeding back into

higher prices for goods or growing asset demand feeding back into higher prices of financial assets, expectations would eventually catch up, potentially causing high inflation in goods or asset prices.

The rational expectations revolution put an end to this line of work. The key criticism is that adaptive expectations feature a particularly unrealistic kind of systematic error. According to what later became known as the Lucas (1976) critique, adaptive expectations do not respond to regime changes. This seems implausible. If a central bank tries to systematically inflate the economy to boost employment, the information that this action is being taken, regardless of past price changes, will promote inflationary expectations. This mechanism was central to accounting for "stagflation" patterns of high unemployment and inflation rates in the 1970s. Likewise, if an economy is stuck in an inflationary spiral but a central bank credibly announces its commitment to end inflation, this information itself, regardless of past price changes, will moderate expected inflation. The backward-looking nature of adaptive expectations and their fixed coefficients do not allow for an immediate response of beliefs to news.

The pure rational expectations solution to this problem is to assume that beliefs are attuned to the key features of the economy, in the specific and extreme sense that expectations are fully dictated by the dynamic model of the economy itself. In the classic formulation of Muth (1961), the rational expectations hypothesis holds that agents know the model that describes the evolution of the economy, observe the shocks that hit it, and based on this information form their expectations as statistically optimal forecasts. These rational forecasts may later turn out to be incorrect, because news can unsettle previous forecasts. But they are correct on average, because they are fully determined by the law that governs the evolution of the economy. A strong prediction follows: under rational expectations, forecast errors cannot be systematically predictable from any information available to the decision-maker at the time the forecast is made.

The rational expectations hypothesis turned out to be one of the most fruitful ideas in the history of economics, forming the foundation of modern macro as an internally coherent and consistent field. But it left several puzzling facts unexplained. In terms of economic outcomes, it had trouble accounting for the slow adjustment of some macroeconomic variables, such as wages or inflation, and for the excess volatility of other variables such as stock prices, interest rates, or home prices. In addition, the assumption that expectations are rational in the sense of not displaying predictable errors was consistently rejected by survey data.

Some early attempts to deal with slow adjustment included theories of rational inattention and information rigidities (Sims 2003; Woodford 2003; Mankiw and Reis 2002; Gabaix 2019), in which agents only partially update their beliefs as new information arrives, due to the cost of absorbing and processing news. Agents are rational, but thinking is costly. Because agents are rational, beliefs are attuned to the model of the economy. Because updating is costly, agents look forward but underreact to news. As a result, the reaction to a shock will be spread out over time, a result that helps a great deal with explaining rigidities in real variables.

The theories of rigid belief changes, however, do not help in a natural way to deal with puzzles related to volatility. In many instances, adjustment to news is strong, and even if it is initially muted, it eventually speeds up as it gets going. In the next section, we show that a resulting pattern of overreaction is indeed present in important macroeconomic data series. Such facts raise two important questions. First, can we build theories of belief formation that can account for excess volatility in expectations, and perhaps even retain some useful features of adaptive expectations, while addressing the fundamental critiques of the ad hoc and backward-looking models raised by Muth and Lucas? Second, can such theories explain expectations data and help account for important macro-finance puzzles? These are the key questions around which our discussion is organized.

Survey Expectations and Predictability of Forecast Errors

The central prediction of the theory of rational expectations is that forecast errors should not be predictable using information known when the forecast was made. A vast body of tests using survey data on the forecasts made by households, professional forecasters, corporate managers, and professional investors nearly universally rejects this prediction.

For example, Souleles (2004) shows that forecast errors in the surveys of consumer confidence and expected inflation from the Michigan Index of Consumer Sentiment do not average out to zero over several decades and are correlated with demographic variables. Greenwood and Shleifer (2014) examine six different data sources for investor forecasts of stock market returns, and find that expectations of future stock returns are too optimistic after stock market booms. Gennaioli, Ma, and Shleifer (2016) study forecasts of earnings growth in a Duke University quarterly survey of chief financial officers, and find that errors can be predicted from past earnings and other factors. Bordalo et al. (2020) consider expectations of 22 macro variables from the Survey of Professional Forecasters and the large-company business economists who participate in the Blue Chip Survey, and find that forecast errors are predictable based on revisions of previous forecasts. Gulen, Ion, and Rossi (2019) find broadly similar results using the same data, along with data from the Institutional Brokers Estimate System (IBES). D'Arienzo (2020) looks at the Blue Chip data on expectations of one-quarter-ahead interest rates on bond yields, and again finds that forecast errors can be predicted based on revisions of previous forecasts. There are many more findings of this kind.

One critique of such findings is that true expectations are unobservable (Prescott 1977), and measured expectations are distorted by a misunderstanding of the survey questions or low incentives for accuracy. This argument is weak for three reasons. First, the evidence overwhelmingly shows that survey expectations are not noise. To begin, elicited beliefs are highly correlated across agents and surveys (for example, as shown in Greenwood and Shleifer 2014). In addition, expressed beliefs typically correlate with economic decisions. In the Gennaioli, Ma, and Shleifer (2016) study,

the expectations of chief financial officers are highly predictive of corporate investment. Giglio et al. (2021) find a correlation between beliefs and portfolio choice in a large survey of sophisticated retail investors with Vanguard. Armona, Fuster, and Zafar (2019) append some questions to the Federal Reserve Bank of New York's Survey of Consumer Expectations, so that randomly selected groups of respondents receive different information, and find that expectations about home price growth have a causal effect on intended investment in housing. In short, the respondents in survey data do actually put their money where their mouths are.

Second, the livelihood of professional stock analysts, macroeconomic forecasters, and corporate managers depends in part on the accuracy of their forecasts. It is hard to maintain that their measured expectations are uninformative about their beliefs. Third, the forecast errors made by different agents often share a systematic overreaction component that cannot be explained by incentives, which differ sharply across agents (say, by demographic or income group, or job).

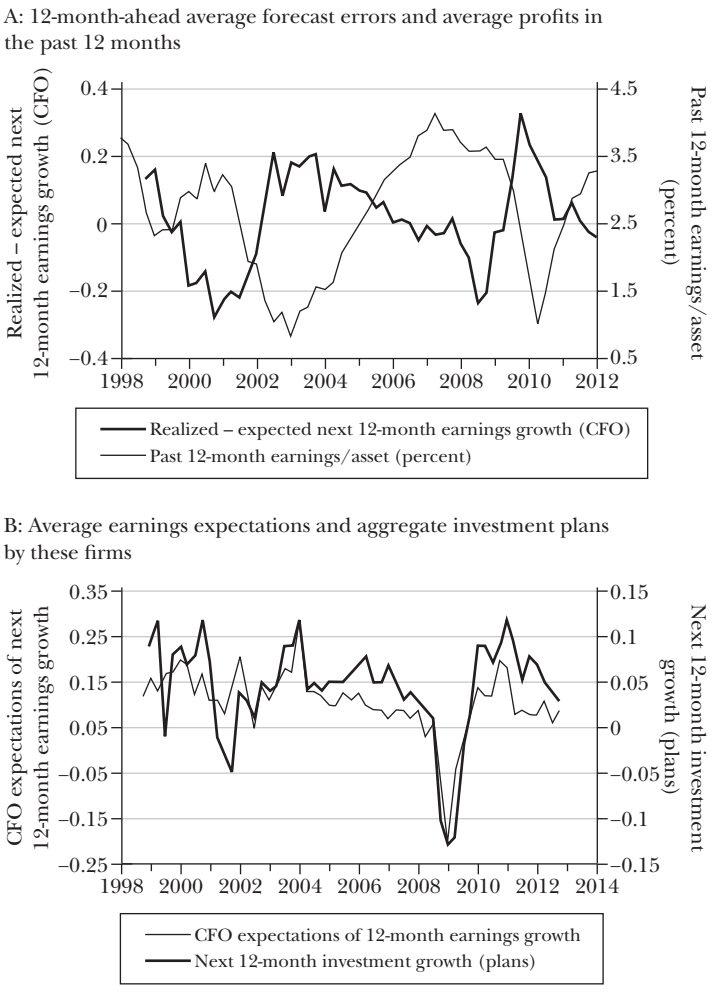
To incorporate survey expectations into macroeconomic analysis, we want to know not just whether forecast errors are systematic, but also whether these errors have meaningful macroeconomic implications. If agents overreact, so they are too optimistic in good times and too pessimistic in bad times, then beliefs are excessively volatile, which translates into excessive volatility in individual decisions. Conversely, if agents underreact so that they are not optimistic enough in good times and not pessimistic enough in bad times, then sluggish belief adjustment translates into sluggish decisions. Different macroeconomic consequences follow in turn.

To detect whether beliefs over- or underreact, two main testing strategies for forecast error predictability have been developed. We describe these tests in turn and present some evidence of how each has been used. The first test correlates the future forecast error, defined as the actual future realization minus the current expectation of a variable, with measures of current conditions. For instance, one can correlate the future error in a manager's earnings growth forecast with the firm's current earnings level.

To see how this works, Figure 1 reports the results obtained when using the expectations of large US-listed companies for their firms' 12-months-ahead earnings growth during the period 1998–2012. As noted earlier, the data is from a Duke University survey of chief financial officers. Panel A plots 12-month-ahead average of forecast errors against average profits in the past 12 months. Panel B plots average earnings expectations and aggregate investment plans by these firms.

Consider the pattern of forecast errors in panel A. If managers' expectations were rational, their future forecast error (thick line) would be uncorrelated with the firm's recent profits (thin line). In contrast, the average future forecast error of managers is strongly negatively correlated with their firms' recent profits: if recent profits have been high (the thin line is high), managerial forecasts are systematically disappointed in the future (the thick line is low). This evidence is indicative of overreaction: good current conditions prompt managers to be too optimistic about the future. Underreaction would predict the opposite: good current conditions would prompt insufficient optimism, which is not the case in the data.

Figure 1
Expectations of Chief Financial Officers, Corporate Profits, and Corporate Investment



Source: Gennaioli, Ma, and Shleifer (2016).
Note: In panel A, the thick line represents aggregate earnings growth in the next 12 months minus aggregate CFO expectations of earnings growth in the next 12 months. The thin line represents aggregate earnings over assets in the past 12 months. Series are linearly detrended. In panel B, the thick line is aggregate planned investment growth in the next 12 months. The thin line is aggregate CFO expectations of next-12-month earnings growth. In both panels, frequency is quarterly.

Overreaction in earnings expectations may shape stock market valuations and firms' investment decisions. Panel B shows that, consistent with this possibility, when the average manager is more optimistic, aggregate investment is higher. Gennaioli, Ma, and Shleifer (2016) show that these patterns are robust to controlling for

aggregate shocks, and that managers' beliefs have a stronger explanatory power for firm-level investment than standard factors such as financing constraints, stock market valuations (as measured by Tobin's *q*), and uncertainty.

The second test for over- versus underreaction of beliefs to news follows from work by Coibion and Gorodnichenko (2012, 2015). Their key innovation is to measure "news" by the extent to which the agent revises the forecast for a fixed future date. The test then consists in assessing whether such forecast revision predicts the agent's future forecast error. This test is conceptually cleaner than the first test, but it is harder to implement because only a few surveys have both a panel structure and the term structure of forecasts necessary to compute forecast revisions.

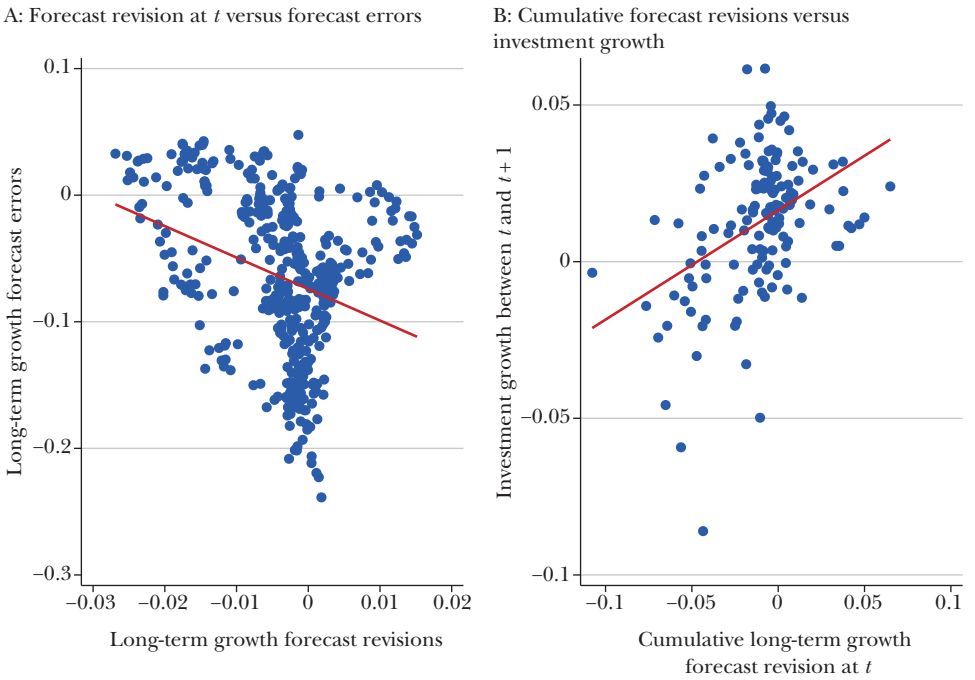
We illustrate the idea of this test using expectations of stock market analysts of long-term earnings growth of listed firms, defined as expected earnings growth over a full business cycle horizon of 3–5 years (La Porta 1996). This data includes forecast revisions. Following Bordalo, Gennaioli, La Porta, and Shleifer (2022), for each firm in the S&P 500 stock index we take the median analyst forecast. We then average these forecasts across firms, obtaining a measure of consensus expectations of aggregate long-term earnings growth. We finally compute revisions in these consensus expectations. Figure 2 presents two plots using this measure of forecast revisions. Panel A plots the five-years-ahead forecast errors in long-term earnings growth against the revision in that variable for the S&P 500 index in the last quarter, over the period 1982–2018. Panel B plots aggregate investment against the current forecast revision for that same earnings growth of S&P 500 firms.

Panel A shows a strong negative correlation between the current forecast revision and the future forecast error. When analysts receive good news (that is, they revise earnings growth forecasts up), their forecasts are systematically disappointed in the future (realized earnings growth is below expectations). This fact is inconsistent with rational expectations, and again points to overreaction: when analysts receive good news, their expectations are revised up excessively and become too optimistic about the future. Underreaction here would predict a positive correlation between forecast revisions and forecast errors, which is not what we see in the data.

Panel B suggests that belief overreaction can have significant economic consequences: current investment growth is strongly positively correlated with the cumulative revision of the long-term growth variable. When the median analyst receives good news (and so do firm managers), aggregate optimism increases and investment rises sharply, perhaps excessively so. Subsequent disappointment of overoptimistic beliefs may cause boom-bust investment cycles.

Coibion and Gorodnichenko originally correlated forecast errors with forecast revisions to assess the theory of rational inattention and information rigidity (Sims 2003; Woodford 2003; Mankiw and Reis 2002). In this theory, the errors of individual analysts should be unpredictable based on their own forecast revisions, but the consensus forecast errors should be positively correlated with the consensus revision. The reason is that individual analysts do not react to information of others, leading to aggregate sluggishness of forecasts. Bordalo et al. (2020) show that,

Figure 2
Forecast Revisions, Forecast Errors, and Investment



Source: Authors' calculations, using the methods in Bordalo et al. (2022).
Note: Panel A plots the five-years-ahead forecast errors in long-term earnings growth against the revision in that variable for the S&P 500 index in the last quarter, over the period 1982–2018. Panel B plots aggregate investment against the cumulative forecast revision for that same earnings growth of S&P 500 firms.

performed at the level of individual forecasters, this test is informative of departures of rational updating. Individual beliefs overreact if the correlation is negative, and underreact if the correlation is positive. Using the Survey of Professional Forecasters and Blue Chip survey data on four-quarters-ahead forecasts for a large set of macroeconomic variables, including measures of economic activity, consumption, investment, and interest rates, they find that, contrary to rational inattention, individual forecasters overreact for most time series. That is, individual analysts do not make optimal use of their own information, but rather overreact, which reveals a deeper problem than rational inattention.

Bordalo et al. (2020) also show that, as long as the information possessed by individual forecasters is limited, which is certainly a realistic assumption, the consensus forecast may appear sluggish even when all individual forecasters overreact. The evidence in panel A of Figure 2 shows that, for stock analysts, overreaction is so strong that it is detectable even at the level of the consensus forecast for the aggregate stock index.

Overall, departures from rational expectations and in particular belief overreaction appear necessary to make sense of the expectations data. Can belief overreaction be formalized and introduced into dynamic macroeconomic analysis? What puzzles in macroeconomics can belief overreaction help address? We answer these questions in the next two sections.

Modeling and Estimating Diagnostic Expectations

In light of the previous discussion, one would like to have a model of belief formation in which expectations capture key features of the structure of the economy, so they have the forward-looking nature that addresses the Lucas (1976) critique. One would also want to have a model in which expectations overreact to information, which is a central fact in survey data.

Over the last several years, we have developed one such model, called diagnostic expectations. This model puts psychology, and in particular selective memory, at center stage (Gennaioli and Shleifer 2010; Bordalo et al. 2016; Bordalo, Gennaioli, and Shleifer 2018). Doing so is key for two reasons. First, while economists in recent decades have mostly relied on preferences to explain challenging facts, psychologists have amassed a substantial body of evidence delineating situations in which beliefs over- or underreact to information (for example, as in Kahneman and Tversky 1972). This evidence is extremely valuable to identify the key properties that a realistic theory of expectation formation should display. Second, memory research has unveiled robust regularities in selective recall (Kahana 2012). Because the information shaping beliefs often comes from memory, these regularities in recall can help build a theory of beliefs from first principles, based on deeper cognitive parameters. The resulting models of expectations can then be more flexible and less ad hoc than old-fashioned adaptive expectations, addressing the Lucas (1976) critique but also accounting for patterns found in survey data.

To motivate the logic of diagnostic expectations, suppose that an agent must assess the future value of a random variable X conditional on data D. The agent has a memory database that contains past realizations of X and of D. Databases may differ across people, due to different experiences, but the main results are already obtained when the database stores the true distribution of events. Critically, when thinking about possible future realizations of X and the data D, the agent automatically and selectively retrieves states X that are most "similar" to the data D compared to other information in the database.[1] The agent who disproportionally samples such distinctive states then overweights their probability in forming expectations.

[1] This assumption reflects the key fact that memory is associative, in the sense that a given event automatically prompts the retrieval of similar events experienced in the past (Kahana 2012). Crucially, similarity between events is measurable, both in terms of frequency of co-occurrence (Tversky 1977; Bordalo, Coffman, et al. 2021) or at a more fundamental level in terms of feature overlap (Bordalo, Conlon, et al. 2021). These measures predict not only subjective similarity assessments but also evidence on recall, probabilistic assessments, and related phenomena.

Suppose for instance that an agent must guess the hair color of a person coming from Ireland, so $X \in \{red, light, dark\}$, and $D = Irish$. As the agent thinks about the possibility that the hair color is $X = red$, many examples of red-haired Irish come to mind. This occurs because Irish people are *more* similar to red hair than other populations, in the sense that red hair is relatively more frequent in Ireland than in the rest of the world. By contrast, as the agent thinks about the possibility that the hair color of an Irish is dark, $X = dark$, few examples of dark-haired Irish come to mind. Indeed, Irish people are much less similar to dark hair than other populations, in the sense that dark hair is relatively *less* frequent in Ireland than in the rest of the world. As a result, even though the dark-haired Irish outnumber the red-haired ones, the agent will oversample from memory the red hair color and overestimate its incidence.[2]

Likewise, when thinking about the health status of someone who tested $D = Positive$ on a medical test, memory oversamples $X = sick$ because this health status is more closely associated with (and hence more similar to) positive as opposed to negative test results. We then overestimate the probability that someone who tested positive has the disease.

This kind of mistake can be especially pronounced when the data points to unlikely and extreme traits. Bordalo et al. (2016) show how this logic accounts for social stereotypes. For instance, people dramatically overstate the prevalence of criminals or terrorists in certain groups, even though an overwhelming majority of any group is honest and peaceful. This bad stereotype is formed automatically when a group contains even a few more criminals than a reference group, which leads to the trait coming to mind more easily.[3] Bordalo, Coffman, Gennaioli, and Shleifer (2019) show that this logic helps explain when and how beliefs about self and others are tainted by gender stereotypes. In a financial setting, this logic explains why investors are likelier to overreact to news that is diagnostic of rare and extreme outcomes (Bordalo et al. 2019; Kwon and Tang 2021).

The model of diagnostic expectations can be used to formalize expectation formation in dynamic settings, as shown formally by Bordalo, Gennaioli, and Shleifer (2018). In that model, forward-looking expectations about an economic variable are based on two components: one component anchored to the rational forecast, and a second component that overweights news received in the most recent few periods.[4] Anchoring to the rational forecast captures the dependence of memory retrieval

[2] Bordalo, Conlon, et al. (2021) present a foundation for stereotypes on the basis of selective recall. Relatively to true frequency, it is harder to think about dark-haired Irish than about red-haired Irish because the former are more similar to other (dark-haired) Europeans. While other Europeans are irrelevant to the task at hand (which is to evaluate the Irish), they are similar to, and interfere with the retrieval of, dark-haired Irish. Red-haired Irish suffer less interference, and therefore are overestimated.

[3] Selective memory generates stereotypes that are not necessarily derogatory; they can be flattering if distinctive traits are good, like a stereotype that "Asian people are good at math."

[4] In formal terms, in such a model an agent's beliefs are captured by the probability density function:

$$f^{\theta}(X|D) \propto f(X|D)\left[\frac{f(X|D)}{f(X|-D)}\right]^{\theta},$$

on the full database, which includes all relevant empirical regularities the agent has experienced. Overweighting of recent news captures disproportionate retrieval of states that are associated with the observed news, which is again shaped by the news events that the agent has experienced in the past. This framework can help unify a great deal of evidence on belief overreaction in macro-financial settings. First, it can produce neglect or overweighting of tail-end downside risk, depending on whether incoming news is good or bad. In good times, good states of the world come to mind and crowd out bad ones, leading to the neglect of tail risk. After a bad shock, bad states of the world come to mind and crowd out the retrieval of good states, leading to exaggerated tail risk.

Second, the model delivers a foundation for extrapolative expectations. As good news causes good outcomes to disproportionally come to mind, and interferes with the retrieval of bad outcomes, the entire distribution of beliefs shifts to the right, causing average excess optimism. The reverse occurs when bad news is received, which causes average excess pessimism. Critically, the extent of extrapolation depends on the data-generating process. For a series with low persistence, news causes a small update in beliefs, because they are less associated with changing future conditions in memory. This prediction is consistent with the evidence from survey data: survey expectations track salient features of the data-generating process. In particular, belief revisions are larger for more persistent series (Bordalo et al. 2020). Unlike for the case of adaptive expectations, updating coefficients are not fixed but rather depend on the underlying reality and have a forward-looking component.

Third, the same mechanism generates systematic reversals in beliefs. Consider the case of an overoptimistic agent. When good news ceases to come in, the agent is no longer cued to oversample good outcomes from memory. As a result, beliefs cool down even in the absence of bad news, causing a sharp reversal that is not driven by bad fundamental news. Diagnostic expectations can generate large movements in beliefs and choices on the basis of small shocks, as well as sudden reversals in beliefs on the basis of past, but not contemporaneous, shocks.

Our diagnostic expectations model is surely not the final formulation, but it offers two advantages relative to alternative theories. First, diagnostic expectations are forward-looking, and respond to changes in the environment using a model of the world. This occurs due to a fundamental feature of human memory: it affects beliefs by causing selective sampling of real-world regularities that are stored in the memory database. As a result, belief distortions depend on the true features of the data-generating process. This aspect is not shared by models in which agents mechanically assume a specific data-generating process, such as one with high

where $f(X|D)$ is the true density, which captures the memory database, and the likelihood ratio captures oversampling of realizations that are relatively more likely given the data D. The strength of oversampling is regulated by $\theta \geq 0$. For $\theta = 0$, beliefs are rational. Bordalo, Gennaioli, and Shleifer (2018) show that when forming beliefs about a Gaussian AR(n) variable, the diagnostic expectation of future value X_{t+1} satisfies $\mathbb{E}_t^{\theta}(X_{t+1}) = \mathbb{E}_t(X_{t+1}) + \theta[\mathbb{E}_t(X_{t+1}) - \mathbb{E}_{t-k}(X_{t+1})]$. In this formula, $\mathbb{E}_t(X_{t+1})$ is the rational forecast, and θ overweights the rational news received in the last k periods. They also estimate the time period k and the magnitude of overstatement θ.

persistence (Angeletos, Huo, and Sastry 2020) or without long-term mean reversion (Fuster, Laibson, and Mendel 2010).

Second, the model of diagnostic expectations can be and has been estimated from empirical data. Critically, its parameters can be compared across different datasets and series/data-generating processes. Several studies have now estimated the parameter controlling the strength of overreaction and found in the survey data on expectations that the reaction to news is about twice what would be warranted under rational expectations (Bordalo et al. 2020; d'Arienzo 2020). These are initial estimates, but they help discipline the ballpark magnitude of overreaction to be used in macroeconomic models.

Belief Overreaction and Macro-Financial Volatility

Overreacting beliefs can help shed light on three central phenomena in finance and macroeconomics: 1) excess stock market volatility, 2) financial crises, 3) regular fluctuations in credit markets and economic activity. They do so in a way that offers hope for a unified approach to economic volatility.

Overreaction and Excess Volatility in the Stock Market

The first, most direct, and perhaps most dramatic evidence of excess volatility comes from the aggregate stock market. Shiller (1981) famously showed that stock prices are much more volatile than warranted by the volatility of future dividends. Campbell and Shiller (1988) further showed that time variation in the price dividend ratio cannot be explained by future dividend growth, but rather by future realized stock returns, which tend to be systematically low after periods in which the price-to-dividend ratio is high.

A growing body of work using survey expectations shows the promise of explaining stock market and more generally financial volatility using overreacting beliefs. One strand of this work is connected to the kind of evidence presented earlier, and argues that stock prices are excessively volatile because beliefs about future dividends or earnings are themselves excessively volatile.[5]

La Porta (1996) first documented that the measure of expected long term earnings growth of earnings (similar to the measure we used in Figure 2) accounts for boom-bust dynamics in the stock price of individual firms: firms that analysts are most optimistic about have lower future stock returns than do firms that analysts are least optimistic about. Bordalo et al. (2019) show that belief overreaction can account for this phenomenon: a firm's high recent earnings growth fuels excess

[5]Another strand of work focuses on extrapolative beliefs about future stock returns. Greenwood and Shleifer (2014) show that investor expectations of one-year-ahead stock returns are too optimistic in good times and too pessimistic in bad times, consistent with overreaction. This may lead to upward price spirals and hence to an overvalued stock market (Barberis et al. 2015). Bordalo et al. (2019) show that controlling for expectations of future stock returns leaves the explanatory power of expectations of fundamentals unaffected.

optimism about its future earnings, which leads to an overvaluation and a future stock price correction as earnings expectations are disappointed. They show that a diagnostic expectations model, with the reaction to news at about twice the rational level, can generate quantitatively realistic boom-bust cycles in expectations and stock prices at the firm level with a realistic process for actual earnings growth.

Can expectations of future fundamentals also account for aggregate stock market volatility? Yes. De la O and Myers (2021) show that time variation in analyst expectations about the market's short-term earnings growth explains a sizable chunk of dividend-price ratio variation. Nagel and Xu (forthcoming) show that a weighted average of past aggregate earnings growth, with weights matching a memory decay rate estimated from inflation expectations (Malmendier and Nagel 2016), correlates with expectations of future earnings growth and low future stock returns. These papers do not, however, focus on systematic errors in measured growth expectations or on their ability to predict future returns. They do not assess whether overreaction drives excess stock market volatility and return predictability.

Bordalo et al. (2022) take up this challenge. They show that, in line with the evidence presented earlier, expectations of aggregate long-term earnings growth indeed overreact, and such overreaction can account for three leading stock market puzzles. First, volatility in expectations of the long-term growth of earnings fully accounts for Shiller's (1981) excess volatility puzzle. Second, overreaction of beliefs about future aggregate earnings growth explains a large share of return predictability in the data. It does so in the aggregate market, accounting for systematically low stock returns after good times and for systematically high stock returns after bad times. But it does so also in the cross section: overreaction of forecasts of aggregate earnings growth accounts for a significant chunk of cross-sectional return spreads typically attributed to risk factors (Fama and French 1992). In this analysis, overreaction of long-term expectations outperforms conventional measures of time-varying risk premia, emerging as a key and parsimonious driver of key stock market puzzles.

Excess volatility has been documented in the bond market as well. Consider the term structure of interest rates, in which long-term interest rates should emerge as an average of short-term rates. Shiller (1979) showed that, from this perspective, long-term interest rates on bonds co-move too much with short-term rates relative to standard benchmarks, a finding he called "excess sensitivity" (Mankiw and Summers 1984; Gürkaynak, Sack, and Swanson 2005). Giglio and Kelly (2018) show that long-term rates are excessively volatile relative to short-term ones, again compared to standard term structure models. They argue that non-rational expectations are needed to explain the evidence. D'Arienzo (2020) directly addresses the role of expectations. Using both survey forecasts from Blue Chip professional forecasters and beliefs extracted from bond prices, he shows that when news arrives, expectations about long-term interest rates overreact compared to those for short-term rates (see also Wang 2021). D'Arienzo (2020) offers a formulation of diagnostic expectations that produces this finding with quantitatively reasonable parametrization. Using a standard term structure model, he shows that such a degree of belief overreaction accounts not only for the bulk of the Giglio and Kelley (2018) excess

volatility puzzle, but also for the excess sensitivity of long-term rates and for bond return predictability (Cochrane and Piazzesi 2009).

In sum, overreaction to news helps account for and unify the evidence of excess volatility and return predictability in the stock and bond markets. Quantitatively, the volatility in measured expectations does a good job accounting for the excess volatility in asset prices.

Overreaction and Financial Crises

Financial crises, defined as episodes of major distress in a country's banking system that are often associated with deep and prolonged recessions, are another leading example of macro-financial volatility. There are two broad rational expectations theories of such crises. In the "bolt from the sky" theories, such crises come as a surprise, such as a large adverse productivity shock, an uncertainty shock (Bloom et al. 2018; Arellano, Bai, and Kehoe 2019), or a "financial shock," which may be a sudden increase in risk aversion or a bank run (Diamond and Dybvig 1983). In the "house of cards" theories, shocks can be small, but hit a financial system that has already been rendered fragile by high leverage. In both cases, the trigger of crises is an exogenous shock, which gets amplified by fire sales, agency problems, or adverse selection (Sufi and Taylor 2021).

Overreacting beliefs suggest a different account, consistent with the informal hypothesis of Minsky (1977) and Kindleberger (1978), as well as with Reinhart and Rogoff (2009). In the boom phase, excessive optimism and neglect of risk fuel asset price bubbles and an overexpansion of credit. When beliefs are systematically disappointed, this causes falling asset values, unsustainable liabilities, fire sales, and panics. As with stock market volatility, a single controlling parameter, the extent of overreaction, accounts for both the boom and the bust.

Large-scale financial crises are sporadic events, many of which occurred a long time ago, so there is no readily available historical data on expectations. This makes it hard to compare theories using measured beliefs. But rational expectations theories make two strong and testable predictions. Under the "bolt from the sky" theories, crises are not predictable. Under "house of cards" theories, crises are predictable with indicators such as high leverage or asset valuations, but markets should show awareness of building up risks since they appreciate the fragility of the system. If in contrast crises are due to belief overreaction, they should be predictable—again, say, based on leverage and valuations—but the pre-crisis period should be associated with euphoria and the neglect of risk (Gennaioli, Shleifer, and Vishny 2015). Data on the predictability of crises as well as on the ex-ante perception of risk can thus distinguish alternative theories.

It is by now well established that the data reject the "bolts from the sky" view: crises are systematically predictable using information on asset prices and quantity of credit.[6] Critically for the current purposes, it also appears that prior to crises,

[6]Borio and Lowe (2002) show that rapid credit and asset price growth predict banking crises in 34 countries between 1970 and 1999. Schularick and Taylor (2012) show that rapid credit expansions in

markets do not exhibit an awareness of heightened risks, as they instead should in the "house of cards" theories. In fact, available evidence suggests that markets exhibit euphoria and dampened risk perceptions before financial crises.

Some of this evidence takes the form of unusually high stock valuation and low credit spreads right before crises.[7] More recent data allow for a closer look at expectations. For the 2007–2008 financial crisis, Jarrow, Mesler, and van Deventer (2006) and Coval, Jurek, and Stafford (2009) show that investors were too optimistic about the returns of securitized assets due to their reliance on incorrect valuation models. Gennaioli and Shleifer (2018) document widespread excessive optimism prior to the Lehman crisis in September 2008, evidenced by homebuyer expectations about future home price growth, investor expectations about the risk of home price declines, and forecasts of economic activity made by both private forecasters and the Federal Reserve. The evidence points to neglect of downside risk in the boom, in line with overreacting expectations.

Overreacting beliefs offer a way to trace the origin of financial crises to a three-stage mechanism reminiscent of Kindleberger (1978). In the first stage, a positive "displacement" such as a technological/financial innovation, or a surge in investor demand, improves an asset's fundamental value. Due to overreaction, expectations become too optimistic, creating an asset price bubble. In the second phase, leverage expands. This effect is amplified by a key byproduct of overreacting beliefs: the neglect of downside tail risk (Gennaioli and Shleifer 2018; Gennaioli, Shleifer, and Vishny 2012, 2013). As a result, even typically risk-averse investors such as banks start to over-expand. In the third phase, beliefs are disappointed, which causes excessive optimism to wane and the asset price bubble to deflate. As risk perception rises, excessive leverage becomes evident, igniting a crisis. In this model, credit spreads are low before the crisis, consistent with the evidence, and the event triggering the crisis is not a negative shock, but the unwinding of the excess optimism created by overreaction to the original, positive shock.[8]

In sum, overreaction to good times and the resulting neglect of downside tail risk help account for financial crises, including the facts that such crises are predictable and begin in what otherwise seem to be good times. Introducing the

---

a sample of 14 developed economies predict financial fragility and bad macroeconomic performance. Mian, Sufi, and Verner (2017) show that growth in household debt predicts low GDP growth in a panel of 30 countries. Most recently, Greenwood et al. (2022) build a predictive index for postwar financial crises using past credit and asset price growth. In a sample of 42 countries over the period 1950–2016, the authors find that a combination of rapid asset price growth and rapid buildup in debt can predict a financial crisis within three years with an over 40 percent probability.

[7] Baron and Xiong (2017) show that, in the run up of bank lending expansions, bank stock returns are unusually high, not low, suggesting neglect of mounting risks. To a similar effect, Krishnamurthy and Muir (2017) shows that crises are typically preceded by unusually low credit spreads.

[8] Recent work has started to model these mechanisms by incorporating diagnostic expectations into a standard model of asset pricing (Bordalo, Gennaioli, Kwon et al. 2021), or into continuous time general equilibrium model of intermediary based asset pricing (Maxted forthcoming; Krishnamurthy and Li 2020; Chodorow-Reich, Guren, and McQuade 2021).

overreaction to news with diagnostic expectations enables otherwise standard dynamic macro models to account for these events.

Business Cycles

The belief formation mechanism may also play a role in regular business cycle fluctuations. Current business cycle research, whether in the New Keynesian or real business cycle model, is almost exclusively built on rational expectations: fluctuations are triggered by demand or supply shocks, which are transmitted via intertemporal substitution and frictions in investment, financing, and price setting. Belief overreaction opens the possibility to connect macroeconomic expansions and recessions to each other via the dynamics of expectations and the systematic winding up and unwinding of optimism.

Business cycles are recurrent events, so the analysis of overreaction can make use of expectations data, which are increasingly available at both aggregate and firm levels. Using post–World War II US data, López-Salido, Stein, and Zakrajšek (2017) find that low credit spreads predict low GDP growth and investment over the next two years.[9] Gulen, Ion, and Rossi (2019) tie these dynamics to expectations data: periods of excess optimism, measured in the ways discussed earlier, are followed by low investment and credit spread reversals.[10]

Can the magnitude of belief overreaction observed in survey data help account for significant business cycle fluctuations? Bordalo, Gennaioli, Shleifer, et al. (2021) address this question by incorporating diagnostic expectations into an otherwise standard real business cycle model with financial frictions. The model is structurally estimated using firm-level data, which crucially includes data on managers' expectations about their firms' profitability. This approach delivers three key results. First, managers' expectations overreact, and the estimated degree of overreaction is similar to that found in other datasets (that is, twice as much as a rational expectations model would predict). Second, the real business cycle model augmented by diagnostic expectations can match successfully untargeted firm-level, as well as sectoral, cycles. Periods when expectations about a firm (or a sector) are overoptimistic, and firm level (sector level) investment is high, are systematically followed by reversals in which i) credit spreads rise, ii) realized bond returns are low, and iii) investment growth is low. Third, the estimated model delivers large increases in aggregate credit spreads, such as the one observed in 2008, from mild reductions in aggregate productivity. The rational version of the same model generates neither systematic boom-bust cycles nor realistic macro-financial volatility without large negative productivity shocks. In this sense, diagnostic expectations offer a

[9] Greenwood and Hanson (2013) show, using US data, that periods in which credit spreads are low, or where a large share of issued bonds are risky, predict disappointing and even negative bond excess returns. Sørensen (2021) shows that periods in which investors accept a low incremental yield for higher default risk in corporate bonds are followed by extremely low returns on risky bonds.

[10] Greenwood and Hanson (2015) document boom-bust cycles in shipbuilding: strong increases in the price of ships lead to excessive investment in shipbuilding and low realized marginal product of investment.

belief-based foundation for the "financial shocks" evident in macro-financial data (Jermann and Quadrini 2012; Gilchrist and Zakrajšek 2012).[11]

This is only the beginning of the systematic assessment of the role of non-rational beliefs in business cycle fluctuations. One important step, for instance, is to connect beliefs with standard mechanisms for demand-driven business cycles such as price rigidity. Bianchi, Ilut, and Saijo (2021) and L'Huillier, Singh, and Yoo (2021) address this question by developing methods to incorporate diagnostic expectations into workhorse New Keynesian models.

In sum, diverse phenomena such as excess stock market volatility, financial crises, and macroeconomic fluctuations may have a common underpinning rooted in overreacting expectations. Two broad messages emerge from the existing work. First, diagnostic expectations enable researchers to incorporate an empirically realistic belief overreaction mechanism into standard dynamic macroeconomic models. Second, the ability of overreaction to produce macro-financial volatility relies on directly measurable expectations.

Alternative Approaches to Macro-Financial Volatility

Economists have grappled with the phenomena of excess financial and economic volatility for decades. Under rational expectations, expectations must on average equal realizations. As a consequence, rational explanations of excess volatility must introduce exogenous variation in preferences or in risk (that is, in required returns for a given degree of risk) to explain the data.

One standard approach, which we call *exotic preferences*, stresses the role of time-varying risk aversion. A prominent example in this class is the idea that preferences are habit-forming, so that the marginal utility of consumption of a representative consumer is very sensitive to even small changes in consumption (Campbell and Cochrane 1999). In good times, when consumption is unusually high, the marginal utility of consumption is very low, and investors accept low expected returns to hold financial instruments to delay consumption. This means, in turn, that valuations are very high. In bad times, when consumption is below trend and the marginal utility of consumption is very high, investors require high returns to hold financial assets, and therefore valuations are very low. The volatility of valuations, and of real variables such as investment, derives from volatility in the marginal utility of consumption.

Another classical approach, which we call *time-varying risk*, introduces high volatility of future risks. In theories of long-run risk (Bansal and Yaron 2004), when investors expect a higher probability of a bad outcome in the distant future, they

[11] These results are due to the fact that diagnostic expectations entail overleveraging in good times, making the economy vulnerable to even small adverse productivity shocks. The explanatory power of the model thus comes from a single parameter controlling overreaction, which is matched using expectations data at the micro level.

avoid risky assets and valuations are low. Fluctuations in expectations about long-run risk can lead to substantial fluctuations in required returns and valuations. A related mechanism focuses on beliefs about the risk of a rare disaster (Barro 2006; Gabaix 2012; Wachter 2013).

These two approaches to resolving the volatility puzzles face closely related problems. First, neither marginal utilities nor long run or rare disaster risk have been systematically measured in the data. These models are driven by unobservables, which can only be inferred from other market outcomes. Second, and more importantly, if we use survey expectations data to evaluate these theories, the evidence rejects both exotic preferences and time-varying risk approaches.

Consider exotic preferences. This approach makes one key prediction about expectations of returns: valuations are high in good times because required (and therefore rationally expected) returns are low. This prediction can be tested using survey evidence on expectations of returns. Greenwood and Shleifer (2014) show, using a variety of investor surveys, that when market valuations are high, expected returns are high, not low. Investors drive up stock prices because they think they will do well, not because they are willing to do poorly. If one takes expectations data seriously, the fundamental premise of exotic preference theories is rejected.

The risk theories do no better. These theories also predict that when risk is high, required (and hence rationally expected) returns should be high. Again, expectations data reject this prediction. Giglio et al. (2021) run a large survey of sophisticated individual investors, and ask them both about their risk perceptions and expectations of stock returns. The paper finds, in a cross section, that investors who expect higher disaster risk also expect lower returns. This of course is exactly the opposite of the prediction of risk theories.

The basic problem of rational models based on exotic preferences or time-varying risk is their inability to account for expectations data and systematic forecast errors, which are indicative of departures from rational updating. A literature on Bayesian learning tries to reconcile the evidence on measured beliefs with rational updating. It shows that systematic forecast errors may arise within a Bayesian framework, provided i) priors are wrong, and ii) learning is slow relative to the frequency of changes in fundamental parameters, such as persistence (Singleton 2021; Farmer, Nakamura, and Steinsson 2021; Timmermann 1993).

The learning approach also stresses the centrality of beliefs and their departure from rationality, which takes the form of wrong priors as opposed to non-Bayesian updating. Despite this similarity with our approach, we see two main problems with the type of learning assumed here. First, the evidence of overreaction is common across variables and datasets. It indicates that recent conditions and news exert an undue influence on beliefs. This seems difficult to reconcile with learning. On the one hand, rational updating would arguably predict dampened reaction to news as agents progressively learn. On the other hand, due to different data-generating processes in different variables and time periods, it would seem that different "wrong priors" would have to be reverse-engineered in order to account for systematic

overreaction across datasets. Overreaction explains a wide range of data by adding just one psychologically well-founded parameter to the rational expectations model.

Diagnostic expectations are one formulation of forward-looking overreaction, and future work should refine this model, in particular bringing in underreaction. Bordalo, Conlon, et al. (2021) show that well-established regularities in human recall, similarity, and interference (Kahana 2012) offer a foundation for the overreaction in diagnostic expectations, but also reconcile it with underreaction to data. The logic of this approach could be used to develop a portable model of belief formation usable in dynamic macroeconomic analysis.

Dynamic macroeconomics, for all its amazing achievements, has resisted taking non-rational expectations seriously. This may be due to a view described by Sargent (2001, paraphrasing Sims 1980), that once we abandon rational expectations, we are in the "wilderness." To us, reality seems to be the reverse: we are in the wilderness if we abandon survey expectations, resorting to unmeasurable mechanisms to account for the data. In contrast, expectations are measurable, understandable from basic psychological principles, disciplined by empirical analysis, and informative about macroeconomics and finance. Departures from rational expectations can be incorporated into models, and the theories can be tested. Unlike in the rational expectations alternatives, theory and evidence go together, and promise a unified view of a great deal of data.

■ *We are deeply grateful to Spencer Kwon, Pierfrancesco Mei, and Johnny Tang for help with the paper.*

References

Angeletos, George-Marios, Zhen Huo, and Karthik A. Sastry. 2020. "Imperfect Macroeconomic Expectations: Evidence and Theory." *NBER Macroeconomics Annual* 35 (1): 1–86.

Armona, Luis, Andreas Fuster, and Basit Zafar. 2019. "Home Price Expectations and Behaviour: Evidence from a Randomized Information Experiment." *Review of Economic Studies* 86 (4): 1371–410.

Arellano, Cristina, Yan Bai, and Patrick J. Kehoe. 2019. "Financial Frictions and Fluctuations in Volatility." *Journal of Political Economy* 127 (5): 2049–103.

Bansal, Ravi, and Amir Yaron. 2004. "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles." *Journal of Finance* 59 (4): 1481–509.

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer. 2015. "X-CAPM: An Extrapolative Capital Asset Pricing Model." *Journal of Financial Economics* 115 (1): 1–24.

Baron, Matthew, and Wei Xiong. 2017. "Credit Expansion and Neglected Crash Risk." *Quarterly Journal of Economics* 132 (2): 713–64.

Barro, Robert. 2006. "Rare Disasters and Asset Markets in the Twentieth Century." *Quarterly Journal of Economics* 121 (3): 823–66.

Bianchi, Francesco, Cosmin Ilut, and Hikaru Saijo. 2021. "Diagnostic Business Cycles." Unpublished.

Bloom, Nicholas, Max Floetotto, Nir Jaimovich, Itay Saporta-Eksten, and Stephen J. Terry. 2018. "Really Uncertain Business Cycles." *Econometrica* 86 (3): 1031–65.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, Frederik Schwerter, and Andrei Shleifer. 2021. "Memory and Representativeness." *Psychological Review* 128 (1): 71–85.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics* 131 (4): 1753–94.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review* 109 (3): 739–73.

Bordalo, Pedro, John J. Conlon, Nicola Gennaioli, Spencer Yongwook Kwon, and Andrei Shleifer. 2021. "Memory and Probability." Unpublished.

Bordalo, Pedro, Nicola Gennaioli, Spencer Yongwook Kwon, and Andrei Shleifer. 2021. "Diagnostic Bubbles." *Journal of Financial Economics* 141 (3): 1060–77.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer. 2019. "Diagnostic Expectations and Stock Returns." *Journal of Finance* 74 (6): 2839–74.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer. 2022. "Belief Overreaction and Stock Market Puzzles." Unpublished.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. 2020. "Overreaction in Macroeconomic Expectations." *American Economic Review* 110 (9): 2748–82.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2018. "Diagnostic Expectations and Credit Cycles." *Journal of Finance* 73 (1): 199–227.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2022. "Replication data for: Overreaction and Diagnostic Expectations in Macroeconomics." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E172721V1.

Bordalo, Pedro, Nicola Gennaioli, Andrei Shleifer, and S. Terry. 2021. "Real Credit Cycles." Unpublished.

Borio, Claudio, and Philip Lowe. 2002. "Assessing the Risk of Banking Crises." *BIS Quarterly Review* 7 (1): 43–54.

Cagan, P. 1956. "The Monetary Dynamics of Hyperinflation." In *Studies in the Quantity Theory of Money*, edited by Milton Friedman, 25–117. Chicago: University of Chicago Press.

Campbell, John Y., and John H. Cochrane. 1999. "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior." *Journal of Political Economy* 107 (2): 205–51.

Campbell, John Y., and Robert J. Shiller. 1988. "Stock Prices, Earnings, and Expected Dividends." *Journal of Finance* 43 (3): 661–76.

Chodorow-Reich, Gabriel, Adam M. Guren, and Timothy J. McQuade. 2021. "The 2000s Housing Cycle with 2020 Hindsight: A Neo-Kindlebergerian View." NBER Working Paper 29140.

Cochrane, John H., and Monika Piazzesi. 2009. "Decomposing the Yield Curve." Unpublished.

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

Coibion, Olivier, and Yuriy Gorodnichenko. 2015. "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review* 105 (8): 2644–78.

Coval, Joshua D., Jakub W. Jurek, and Erik Stafford. 2009. "Economic Catastrophe Bonds." *American Economic Review* 99 (3): 628-66.

d'Arienzo, Daniele. 2020. "Maturity Increasing Overreaction and Bond Market Puzzles." Unpublished.

De la O, Ricardo, and Sean Myers. 2021. "Subjective Cash Flow and Discount Rate Expectations." *Journal of Finance* 76 (3): 1339–87.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91 (3): 401–19.

Fama, Eugene F., and Kenneth R. French. 1992. "The Cross-Section of Expected Stock Returns." *Journal of Finance* 47 (2): 427–65.

Farmer, Leland, Emi Nakamura, and Jón Steinsson. 2022. "Learning about the Long Run." Unpublished.

Fuster, Andreas, David Laibson, and Brock Mendel. 2010. "Natural Expectations and Macroeconomic Fluctuations." *Journal of Economic Perspectives* 24 (4): 67–84.

Gabaix, Xavier. 2012. "Variable Rare Disasters: An Exactly Solved Framework for Ten Puzzles in Macro-Finance." *Quarterly Journal of Economics* 127 (2): 645–700.

Gabaix, Xavier. 2019. "Behavioral inattention." In *Handbook of Behavioral Economics: Applications and Foundations 2*, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, 261–343. Amsterdam: North-Holland.

Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer. 2016. "Expectations and Investment." *NBER Macroeconomics Annual* 30 (1): 379–431.

Gennaioli, Nicola, Andrei Shleifer, and Robert W. Vishny. 2012. "Neglected Risks, Financial Innovation, and Financial Fragility." *Journal of Financial Economics* 104 (3): 452–68.

Gennaioli, Nicola, Andrei Shleifer, and Robert W. Vishny. 2013. "A Model of Shadow Banking." *Journal of Finance* 68 (4): 1331–63.

Gennaioli, Nicola, Andrei Shleifer, and Robert W. Vishny. 2015. "Neglected Risks: The Psychology of Financial Crises." *American Economic Review* 105 (5): 310–14.

Gennaioli, Nicola, and Andrei Shleifer. 2010. "What Comes to Mind." *Quarterly Journal of Economics* 125 (4): 1399–433.

Gennaioli, Nicola, and Andrei Shleifer. 2018. *A Crisis of Beliefs*. Princeton: Princeton University Press.

Giglio, Stefano, and Bryan Kelly. 2018. "Excess Volatility: Beyond Discount Rates." *Quarterly Journal of Economics* 133 (1): 71–127.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus. 2021. "Five Facts about Beliefs and Portfolios." *American Economic Review* 111 (5): 1481–522.

Gilchrist, Simon, and Egon Zakrajšek. 2012. "Credit Spreads and Business Cycle Fluctuations." *American Economic Review* 102 (4): 1692–720.

Greenwood, Robin, and Samuel G. Hanson. 2013. "Issuer Quality and Corporate Bond Returns." *Review of Financial Studies* 26 (6): 1483–525.

Greenwood, Robin, and Samuel G. Hanson. 2015. "Waves in Ship Prices and Investment." *Quarterly Journal of Economics* 130 (1): 55–109.

Greenwood, Robin, and Andrei Shleifer. 2014. "Expectations of Returns and Expected Returns." *Review of Financial Studies* 27 (3): 714–46.

Greenwood, Robin, Samuel G. Hanson, Andrei Shleifer, and Jakob Ahm Sørensen. 2022. "Predictable Financial Crises." *Journal of Finance* 77 (2): 863–921.

Gulen, Huseyin, Mihai Ion, and Stefano Rossi. 2019. "Credit Cycles, Expectations, and Corporate Investment." CEPR Discussion Paper 13679.

Gürkaynak, Refet S., Brian Sack, and Eric Swanson. 2005. "The Sensitivity of Long-Term Interest Rates to Economic News: Evidence and Implications for Macroeconomic Models." *American Economic Review* 95 (1): 425–36.

Jarrow, R., M. Mesler, and D.R. van Deventer. 2006. *Default Probabilities Technical Report*, Version 4.1. Honolulu: Kamakura Risk Information Services, Kamakura Corporation.

Jermann, Urban, and Vincenzo Quadrini. 2012. "Macroeconomic Effects of Financial Shocks." *American Economic Review* 102 (1): 238–71.

Kahana, Michael Jacob. 2012. *Foundations of Human Memory*. New York: Oxford University Press.

Kahneman, Daniel, and Amos Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3 (3): 430–54.

Kindleberger, Charles P. 1978. *Manias, Panics and Crashes: A History of Financial Crisis*. New York: Basic Books.

Krishnamurthy, Arvind, and Tyler Muir. 2017. "How Credit Cycles across a Financial Crisis." NBER Working Paper 23850.

Krishnamurthy, Arvind, and Wenhao Li. 2020. "Dissecting Mechanisms of Financial Crises: Intermediation and Sentiment." NBER Working Paper 27088.

Kwon, S.Y., and J. Tang. 2021. "Extreme Events and Overreaction to News." Unpublished.

La Porta, Rafael. 1996. "Expectations and the Cross- Section of Stock Returns." *Journal of Finance* 51 (5): 1715–42.

L'Huillier, Jean Paul, Sanjay R. Singh, and Donghoon Yoo. 2021. "Diagnostic Expectations and Macroeconomic Volatility." Unpublished.

López-Salido, David, Jeremy C. Stein, and Egon Zakrajšek. 2017. "Credit-Market Sentiment and the Business Cycle." *Quarterly Journal of Economics* 132 (3): 1373–426.

Lucas, Robert E. 1976. "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy* 1: 19–46.

Lucas, Robert E., and Thomas J. Sargent, eds. 1981. *Rational Expectations and Econometric Practice*, Vol. 2. Minneapolis: University of Minnesota Press.

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

Mankiw, N. Gregory, and Lawrence H. Summers. 1984. "Do Long-Term Interest Rates Overreact to

Short-Term Interest Rates?" *Brookings Papers on Economic Activity* 15 (1): 223–48.
Mankiw, N. Gregory, and Ricardo Reis. 2002. "Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *Quarterly Journal of Economics* 117 (4): 1295–328.
Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers. 2003. "Disagreement about Inflation Expectations." *NBER Macroeconomics Annual* 18: 209–48.
Maxted, Peter. Forthcoming. "A Macro-Finance Model with Sentiment." *Review of Economics Studies***.**
Mian, Atif, Amir Sufi, and Emil Verner. 2017. "Household Debt and Business Cycles Worldwide." *Quarterly Journal of Economics* 132 (4): 1755–817.
Minsky, Hyman P. 1977. "The Financial Instability Hypothesis: An Interpretation of Keynes and an Alternative to 'Standard' Theory." *Challenge* 20 (1): 20–7.
Muth, John F. 1961. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (3): 315–35.
Nagel, Stefan, and Zhengyang Xu. Forthcoming. "Asset Pricing with Fading Memory." *Review of Financial Studies.*
National Bureau of Economic Research. 1960. *The Quality and Economic Significance of Anticipations Data.* Princeton: Princeton University Press.
Prescott, Edward C. 1977. "Should Control Theory Be Used for Economic Stabilization?" *Carnegie-Rochester Conference Series on Public Policy* 7 (1): 13–38.
Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. *This Time Is Different.* Princeton: Princeton University Press.
Sargent, Thomas J. 2001. *The Conquest of American Inflation.* Princeton: Princeton University Press.
Schularick, Moritz, and Alan M. Taylor. 2012. "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870–2008." *American Economic Review* 102 (2): 1029–61.
Shiller, Robert J. 1979. "The Volatility of Long-Term Interest Rates and Expectations Models of the Term Structure." *Journal of Political Economy* 87 (6): 1190–219.
Shiller, Robert J. 1981. "Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends?" *American Economic Review* 71 (3): 421–36.
Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48 (1): 1–48.
Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–90.
Singleton, Kenneth J. 2021. "Presidential Address: How Much 'Rationality' Is There in Bond-Market Risk Premiums?" *Journal of Finance* 76 (4): 1611–54.
Sørensen, J. 2021. "Risk Neglect in the Corporate Bond Market." Unpublished.
Souleles, Nicholas S. 2004. "Expectations, Heterogeneous Forecast Errors, and Consumption: Micro Evidence from the Michigan Consumer Sentiment Surveys." *Journal of Money, Credit and Banking* 36 (1): 39–72.
Sufi, Amir, and Alan M. Taylor. 2021. "Financial Crises: A Survey." NBER Working Paper 29155.
Timmermann, Allan. 1993. "How Learning in Financial Markets Generates Excess Volatility and Predictability in Stock Prices." *Quarterly Journal of Economics* 108 (4): 1135–45.
Tversky, Amos. 1977. "Features of Similarity." *Psychological Review* 84 (4): 327–52.
Vissing-Jorgensen, Annette. 2003. "Perspectives on Behavioral Finance: Does 'Irrationality' Disappear with Wealth? Evidence from Expectations and Actions." *NBER Macroeconomics Annual* 18: 139–208.
Wachter, J.A. 2013. "Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?" *Journal of Finance* 68 (3): 987–1035.
Wang, C. 2021. "Under- and Overreaction in Yield Curve Expectations." Unpublished.
Woodford, Michael. 2003. "Imperfect Common Knowledge and the Effects of Monetary Policy." In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, edited by Philippe Aghion et al., 25–58. Princeton: Princeton University Press.

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 245–260

Retrospectives
On the Evolution of the Rules versus Discretion Debate in Monetary Policy

Harris Dellas and George S. Tavlas

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please contact either Beatrice Cherrier, CNRS & CREST, ENSAE-Ecole Polytechnique (beatrice.cherrier@gmail.com) or Joseph Persky, University of Illinois at Chicago (jpersky@uic.edu).

Introduction

Monetary policy is said to follow a rule when it satisfies the promises it has made about how it will react in the future to observable economic developments. For instance, a typical pure inflation targeting rule implies that the policy interest rate is increased by more than one to one with the excess of inflation over the inflation target. Similarly, in the celebrated Taylor rule, the policy rate responds in a predetermined way to excesses of inflation and/or unemployment relative to target. In open economies, a common rule has the central bank adjusting its instrument to maintain a fixed exchange rate. These rules are called "activist," because they specify an action to be taken.

■ *Harris Dellas is Professor and Director, Institute of Political Economy, University of Bern, Bern, Switzerland, and Research Fellow, Centre for Economic Policy Research (CEPR), London, United Kingdom. George S. Tavlas is Alternate to the Governor of the Bank of Greece on the European Central Bank (ECB) Governing Council and Distinguished Visiting Fellow, Hoover Institution, Stanford, California. Tavlas is the corresponding author at gtavlas@bankofgreece.gr.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.245.

A policy rule does not necessarily prescribe a response to developments in the economy; it may involve exogenous behavior. A rule that the money supply should grow at a constant rate is an example. Exogenous rules are called "passive."

Alternatively, a central bank may make promises of actions that it does not always keep. For instance, it may announce that it will raise interest rates if inflation exceeds its target, but then fail to do so. Not being bound by promises is the key characteristic of discretionary policy. An example seems to be US monetary policy during the 1970s when the Federal Reserve repeatedly promised to tighten policy to tame inflation, but did not follow through when faced with higher-than-desired unemployment rates.

What are the implications of following a rule (keeping promises) rather than exercising discretion in responding to changing economic circumstances (not keeping promises)? Does one policy regime systematically produce better outcomes than the other? How and why? We present the views of the protagonists in systematic debates about rules vs. discretion that arose during three major historical episodes. Unsurprisingly, all three debates sprang from events associated with economic turbulence and poor monetary policy performance. Most readers will be familiar with the first two occasions: the stagflation of the 1970s and the Great Depression of the early 1930s. The third case, the Currency School versus Banking School debates that arose in the United Kingdom in the 1820s in a deflationary period following the Napoleonic wars, is included because it is commonly cited as one of the first explicit "rules versus discretion" debates over monetary policy.

We highlight the main features and results of the debates that emerged in the three episodes and we identify their most important differences. The first and second historical debates emphasized conflicts of interest between different agents in the economy as the driver of discretion and identified greater uncertainty as its main inefficiency. They also contained discussions about the macroeconomic properties of some simple rules, whether passive or activist.[1] In contrast, the literature that emerged from the modern debate did not view discretion as a consequence of a disharmony between public and private interests. It established that discretion is inefficient even when it does not lead to greater uncertainty. It emphasized the role played by rules for expectations management (such as forward guidance). It also showed that simple rules—such as the Taylor rule—could deliver outcomes that were close to the optimal ones.

[1] In practice, it may be difficult to distinguish between an activist rule and discretion. For example, although the nineteenth-century Currency School versus Banking School debate is often cited as a case of rules versus discretion, we argue below that it is more accurately interpreted as a dispute over the appropriate degree of activism in the policy rule under a gold standard.

The 1970s Stagflation and Monetary Rules

The 1970s represent a great failure of macroeconomic policy in the United States (and many other countries), with high rates of both inflation and unemployment, along with considerable macroeconomic volatility. There is a widespread consensus that "the dominant inflation impulse came from monetary policy" (Meltzer 2010, p. 844). The monetary policy literature that emerged from this Great Inflation episode identified policy discretion in the face of unsustainable targets for unemployment amid adverse shocks as the main culprit. Moreover, it showed how the policymakers' failure to keep promises and to subdue inflation fuelled high inflation expectations, which, in turn, worsened policy trade-offs, contributing to higher current inflation by inducing economic agents to shift spending toward the present.

The rational expectations revolution of the 1970s provided tools for studying how policy conduct impacts on expectations, how expectations matter for policy now and in the future, and how both policies and expectations matter for economic decisions. In general, current policies depend on current economic conditions. But current economic conditions depend on agents' current decisions, which in turn depend on expectations of future policy. Future policy depends on future economic conditions which are partly shaped by current economic decisions made on the basis of current expectations of future policies. Expectations thus create a bi-directional link between present and future policies.

This link is at the heart of the concept of time consistency. A policy is time-consistent if what it prescribes at time T for all future times $T + t$ remains the best action when time $T + t$ arrives. A policy is time-inconsistent if it is no longer represents the best action to take when $T + t$ arrives. Discretionary policies are time-consistent, and thus feasible, because they represent the best action that a policymaker can take in any period irrespective of what policy choices were made in the past or what policy decisions are expected to be made in the future. As discussed below, optimal policies are, in general, time-inconsistent. This is due to the fact that time consistency requires future policymakers to behave in a way that is consistent with the previously-formed expectations of policy. But *in the absence of policy commitment*, there is nothing that compels policymakers to honor previously formed expectations: they may elect not to do so and, instead, pursue discretionary policies that are inconsistent with those expectations. Naturally, if agents have some notion of the structure of the economy and are endowed with rational expectations, they will be able to infer such behavior and adjust expectations accordingly. Herein lies the problem. Current expectations about the future shape current economic decisions and, thus, matter for current optimal policy. If the expectations are not the "right" ones, then current policy cannot be the right one either. One way to manage current expectations about the future so that they are "right" is by adhering to a policy rule.

The modern literature on the problem of time-inconsistency started with the seminal contribution of Kydland and Prescott (1977). Barro and Gordon (1983)

followed in their footsteps and provided a positive theory of inflation according to which policymakers' use of discretionary policies to attain short-term unemployment objectives could account for the experience of the 1970s. These works rely on rational expectations models of the Phillips curve that, given expectations, involve a short run trade-off between inflation and unemployment. They showed how attempts to exploit this trade-off via discretionary policy could eliminate the trade-off and produce excessive inflation without any corresponding benefit on unemployment. The key to this result lies in the property of rational expectations, according to which the public correctly anticipates the propensity of policymakers for positive inflation surprises and adjusts its expectations correspondingly; in the end (in equilibrium), there are no inflation surprises and employment gains but policymakers are left with a high level of anticipated and actual inflation.

It should be stressed, however, that discretionary policy in any period does not necessarily represent myopic or misguided behavior *from the point of view of that period*. As Kydland and Prescott (1977, p. 481, italics in original) put it: "The reason that such policies are suboptimal is not due to myopia . . . Rather, the suboptimality arises because there is no mechanism to induce *future* policymakers to take into consideration the effect of their policy, via the expectations mechanism, upon *current* decisions of agents."

Moreover, the issue is not that economic agents are uncertain about what policymakers will do in the future (a key theme in the earlier literatures); on the contrary, agents know that policymakers under discretion will do the "wrong" thing, that is, given expectations, policymakers will attempt to generate more inflation in order to decrease unemployment.

Calvo (1978) reformulated the Kydland and Prescott (1977) model in a way that has become the standard approach in the study of optimal policy. A key innovation of this approach is that optimal policy is determined by having the government maximize the utility of the representative agent. Note that this formulation implies that private agents and the policymakers share the same objective, which made it transparent that the time-inconsistency-infeasibility of the optimal plans does not rest on any disharmony or conflict between the objectives of private agents and the government. Another innovation concerns the demonstration that, for time inconsistency to be an issue for optimal policy, there must be some distortion in the economy that gives rise to a policy trade-off, which, in turn, motivates the choice of discretionary actions. Some form of this trade-off is present in all models of time inconsistency, but is often not transparent.[2]

[2] In Calvo's (1978) model, the implementation of optimal policy requires taxes to be used to facilitate the issuance or absorption of money. Had lump-sum taxes been available to carry out this function, the distortionary inflation tax would have been an inferior source of tax revenue, there would be no need for seigniorage surprises, and optimal inflation policy would have been time consistent. In the literature that followed Barro and Gordon (1983), this trade-off is usually generated from the assumption that the natural level of output is inefficiently low. Were it not for this assumption, optimal policy would be time-consistent and there would be no inflation bias.

For these reasons, a built-in "inflation bias" is a key disadvantage of discretionary monetary policy. Once random disturbances are admitted to the economy, though, discretionary policy has an additional disadvantage of a suboptimal response to certain unanticipated shocks (Woodford 2003, Chapter 7.2; Clerc, Dellas, and Loisel 2011). The main result here is that if the optimal policy response can be delivered in a single time period, as is theoretically the case with demand shocks, then either discretionary or rule-based policies can be optimal. But for other shocks, like supply shocks (or cost-push) shocks, the optimal response is delivered in installments, spread out over several time periods, thus producing a smoother output gap (Galí 2008, Chapter 5.2.2). In this case, discretionary policy will be inferior to an optimal rules-based policy, because, due to distrust, it cannot rely on expectations management to smooth the total response over time and to deliver the policy in credible "installments."[3] This is called the "stabilization bias" of discretionary policy.

How can optimal monetary policy free itself from the discretion temptation and become time consistent? Kydland and Prescott (1977) suggested, without elaborating, the use of easily understood rules. A voluminous subsequent literature has been exploring the forms and properties of alternative rules, differentiating them according to various criteria: state-contingent or not, purely forward-looking or history-dependent, flexible or rigid, optimal or suboptimal (but still performing better than discretion), targeting an outcome variable or a policy instrument variable, passive or activist, and so on. A discussion of alternative monetary policy rules can be found in Woodford (2003, Chapters 7 and 8).

A key distinction is between passive and activist rules. Under a passive rule, such as a rule for steady growth of the money supply or an exchange rate target, policymakers are obliged to follow the same course of action in all circumstances. Under an activist rule, as in the case of a strict inflation targeting or a Taylor type of rule that specifies a response of monetary policy to inflation and unemployment rates, policymakers can respond to different circumstances in pre-determined ways. Activist rules, in general, tend to outperform passive rules. Moreover, even relatively simple activist rules like a version of the Taylor rule may be close to optimal, at least in some circumstances. This robust finding survives the presence of imperfect information in the conduct of policy (limited knowledge of the structure of the economy and of the effects of policies, due, say, to time lags).

Other than explicit rules, might other mechanisms both discourage discretionary behavior and, at the same time, allow for welfare-improving policy "flexibility"? Walsh (2010) contains a comprehensive treatment of the possibilities. The main mechanisms concern reputation (credibility) building, appointing

[3]This point relates to the observation made earlier that in the absence of a policy trade-off, optimal policy is time-consistent. Another way of expressing this same conceptual point is that shocks that do not confront optimal policy with a trade-off between inflation and output, like "demand" shocks, can be dealt with equally efficiently by both discretionary and optimal policies. However, for shocks that do give rise to a trade-off (such as "cost-push" shocks), discretionary policy is inferior to an optimal rules-based policy.

policymakers who are "conservative" in the sense that they are unlikely to alter policy away from earlier promises, offering suitable contracts to policymakers, and undertaking institutional reforms that set and enforce an appropriate mandate. The reputation sub-literature has made heavy use of game theory with incomplete information, while the contracting sub-literature has relied on the principal-agent approach. Both sub-literatures have thus introduced up-to-date micro-economic tools in the analysis of optimal policy.

The rigorous study of optimal policy in dynamic settings led to new insights and had a big impact on the design of monetary (and fiscal) policy in the modern era. The modern literature emphasized that managing inflation expectations properly—that is, creating the correct sort of expectations about future policy—is an essential element of optimal policymaking. Thus, the key failure of discretionary policymaking is precisely making decisions at each time that tend to ignore the extent to which expectations are influenced by, and, in turn, influence, government policy. It also argued that the reasons behind discretion need not require conflicts of interest or uncertainty about the actions of policymakers (which, as we shall see, was a main theme in the earlier literature). It showed that optimal policies will in general be time-inconsistent, and it prescribed the use of policy rules as a solution. Ultimately, these insights led to a fundamental rethinking of the implementation of monetary policy in favor of rules-based policies, with the main examples being the widespread adoption of central-bank independence and inflation-targeting rules (for discussion, see Taylor 2017).

The Great Depression as a Failure of Central Bank Discretion

The legislation creating the US Federal Reserve was signed in December 1913. After dealing with wartime finance when World War I started in July 1914, the newborn Federal Reserve in 1918 began to conduct monetary policy via discount rate and open-market operations with the aim of achieving three objectives that Meltzer (2002, pp. 261–262) characterized as "incompatible": the restoration of the international gold standard, the prevention of inflation, and the mitigation of business fluctuations. By the end of the 1920s, a fourth objective, dampening of stock-market speculation, had been added to the list. The Fed acted in a highly unsystematic and unpredictable fashion during the 1920s, pursuing inconsistent objectives and changing them from year-to-year, fostering uncertainty about monetary policy. Friedman and Schwartz (1963, p. 297) characterized monetary policy during the 1920s as follows: "Inevitably, in the absence of any single well-defined statutory objective, conflicts developed between discretionary objectives of monetary policy. The two most important arose out of the re-establishment of the gold standard abroad and the emergence of the bull market in stocks." We will not seek here to rehearse the links from monetary policy and the banking system to the Great Depression: suffice it to say that by the early 1930s, it was clear that macroeconomic policy had gone badly wrong.

In November 1933, University of Chicago economist Henry Simons wrote and circulated a 27-page unpublished memorandum, "Banking and Currency Reform," in which he advanced a set of policies for combatting the Great Depression and avoiding such episodes in the future.[4] A few years later, Simons (1936) published "Rules versus Authorities in Monetary Policy," which expanded on the policy proposals set forward in the 1933 memorandum. In these papers, Simons made two lasting contributions. First, he evaluated the benefits and costs of *alternative* monetary-policy rules: a steady percentage increase in the quantity of money, a stable price level, a fixed quantity of money, a fixed quantity of money per capita, a moderately declining price level; and the gold standard. Second, Simons (1936, pp. 163–64) provided *criteria* to be used in assessing the merits of the alternative rules: freedom from political interference, simplicity (that is, ease of communicating the rule), definiteness, compatibility with fiscal discipline, and the absence of judgement in the administration of the rule.[5]

In his 1933 memorandum, Simons favored a passive rule that fixes the quantity of money because, he argued, it would both bind the authorities to a policy instrument—the money stock—and deliver an objective—economic stability. By 1936, however, Simons had come to favor an activist rule defined by a focus on stabilizing the price level, although he recognized that such a rule could allow the policymakers to exercise discretion in the choice and use of policy instruments. The limitations of a fixed money quantity rule, Simons (1936, p. 171) explained, "have to do mainly with the unfortunate character of our financial structure—with the abundance of what we may call 'near moneys'—with the difficulty of defining money in such a manner as to give practical significance to the conception of quantity."

The single most important attribute of a policy rule, according to Simons (1936, p. 161), was its ability to minimize policy uncertainty: "An enterprise system cannot function effectively in the face of extreme uncertainty as to the action of the monetary authorities or, for that matter, as to monetary legislation. We must avoid a situation where every business venture becomes largely a speculation on the future of monetary policy." He believed that, apart from the gold standard rule, any one of the other rules would be preferable to a discretionary regime because the existence of a rule would minimize uncertainty.

In the 1940s and early 1950s, Lloyd Mints, Simons's colleague at the University of Chicago, followed Simons on the rules-versus-discretion issue. In *Monetary Policy for a Competitive Society*, Mints (1950, p. 8) characterized the Fed's discretionary policies in the 1920s as follows:

> During the 1920's this belief [in the power of central-bank policy] was greatly strengthened, and what were held to be the goals of central-bank action were

[4]The memorandum was widely distributed (Phillips 1995, p. 49). In drafting the memorandum, Simons received substantial input from Aaron Director (Tavlas forthcoming, chap. 2).

[5]During the 1920s, economists such as Irving Fisher had argued that the Fed should pursue the single objective of price-level stabilization, which amounted to a policy rule. But those economists did not cast their advocacy of price-level stabilization in the context of a preference for rules over discretion; nor did they assess alternative policy rules (Tavlas 2022).

> more explicitly formulated. The most unfortunate aspect of this development was the general belief that the central bank should be given wide discretionary powers to take whatever action seemed to it wise in given circumstances. The Federal Reserve System was created and was operated (and still is) in accordance with this point of view.

Mints (1950, p. 46, fn. 5) explicitly attributed the "tragic failure" of monetary policy during the Great Depression to discretionary management—and not to the particular individuals in power: "I intend that my criticisms of the Reserve System shall be unambiguous and largely adverse; but I do not mean to imply that another group of men, under the same conditions and operating with the same grant of discretionary power, would have done better. It is to discretionary monetary authorities, that I object."

Like Simons, Mints (1950) considered alternative policy rules and provided criteria to assess their merits, coming out in favor of a price-level-stabilization rule, and, also like Simons, Mints emphasized that the most important attribute of a rule is its ability to reduce policy uncertainty. Simons and Mints both identified management of expectations as the key advantage of rules over discretion. They maintained that a policy rule would help stabilize expectations by reducing policy uncertainty, thus helping to dampen economic fluctuations. Mints (1946, p. 60, italics added) even argued that, under a rule that stabilizes the price level, "aggregate demand could be quickly restored by monetary-fiscal measures, *if not by mere expectations of such measures*, and thus nothing more than a minor recession in business activity need ever arise."

However, Mints (1950) also added several new elements to these arguments. First, while Mints supported an activist rules policy (targeting the price level), he was distrustful of policies that "would require [the central bank] . . . [to] be able to forecast economic conditions with at least a fair degree of accuracy and for a considerable period of time in advance," an ability that Mints (1945, p. 279) thought that central banks did not possess. In this regard, Mints anticipated some of the arguments in the modern debate questioning the wisdom of adopting forward-looking rules, such as a Taylor-type rule that responds to deviation of inflation forecasts from some target level (Galí 2008, Chapters 3.1.3 and 4). Second, Mints brought attention to the fact that monetary-policy actions were subject to long and variable time lags, which made it difficult to predict the effects of those actions. Mints argued that the existence of lags would accentuate the uncertainty created by discretionary policy.[6]

By the 1960s and 1970s, Milton Friedman became the best-known proponent of a monetary policy rule: in particular, from 1956 onward, Friedman (1960, p. 91) favored a passive rule under which the money supply would grow at a rate between 3 to 5 percent per year to attain economic stability and "a roughly stable price level." He cited a number of anti-discretion arguments, including: 1) discretion has "meant continued and unpredictable shifts in the immediate guides to policy and in

[6]Mints (1950, p. 138, n. 8) credited Friedman for identifying the problem created by the *variability* of lags in accentuating uncertainty. Earlier, Mints (1945) had identified the problem created by the length of lags. For an assessment of Mints's contributions to monetary economics, see Dellas and Tavlas (2021).

the content of policy as the persons and attitudes dominating the authorities have changed" over time; 2) discretion exposed the authorities "to political and economic pressures and to the deceptive effects of short-lived ideas of events and opinions"; and 3) reliance on discretion in pursuing general goals "meant also the absence of any criteria for judging performance." Moreover, Friedman (1953, pp. 129–131) believed that limited knowledge about the lags in monetary policy could make discretionary policies destabilizing (for discussion, see Nelson 2020, p. 301). In short, concerns about imperfect information and conflicting political influences led him to reject activist rules in favor of a passive rule of targeting the money supply.

This earlier literature on the choice between rules and discretion had some similarities with the modern literature. Both were motivated by a failure of monetary policy on a grand scale. Both recognized that the critical factor underlying the inefficiency of discretion was its inability to fruitfully manage expectations. Both considered alternative rules and were cognizant of the trade-offs involved in activist versus passive rules.

However, the pre-1970 and modern literature on this topic have some key differences. In the earlier literature, policy discretion was viewed as synonymous with unpredictable policy, and expectations destabilized by this uncertainty played the key role in demonstrating the inefficiency of discretion. Policy uncertainty plays a less prominent role in the modern literature, which focuses instead on problems of time-consistency and built-in inflation bias—problems that would exist even with a predictable discretionary policy. In addition, while the earlier literature opposed unfettered discretion, it also tended to prefer passive rules, like targeting the money supply, or narrow activist ones, like targeting the price level. In contrast, the modern literature has considered broad, activist, and informationally demanding rules.

The Currency School versus Banking School Controversy

In 1793, the British government declared war on revolutionary France, precipitating a drain of gold from the British banking system. In February 1797, the Bank of England—then a private institution at the center of the British financial system—reported to the government that its gold reserves had fallen to such a low level that it would not be able to remain open. Thus, the Bank requested, and the government approved, a prohibition of the Bank's exchanging its notes for specie. From 1797 to the end of the Napoleonic Wars in 1815, there was a widespread perception that British prices had risen sharply, based on the premium of bullion over the face value of paper currency and the discount of sterling against other currencies relative to the metallic parities of the pound and those currencies.[7] With the end of the Napoleonic Wars in 1815, the British economy entered a deflationary and recession-plagued phase that persisted through the 1820s.

[7]The suspension of specie payments set the stage for the Bullionist controversy that took place in the first two decades of the nineteenth century. The key issue addressed was the following: what caused the

Following parliamentary debates, convertibility was reinstituted in 1821. Three severe financial crises—in 1825, 1836, and 1839—marked the following 20 years. The crises took the form of bank runs, as holders of bank notes and banks' depositors sought refuge by converting their wealth to the safety of gold.

The 1825 crisis, in particular, marked the beginning of the debate that would last for several decades between what became known as the Currency School and the Banking School. Members of the Currency School included Samuel Jones Loyd (later Lord Overstone), Robert Torrens, and George W. Norman. Members of the Banking School included John Fullarton, John Stuart Mill, Thomas Tooke, and James Wilson. The debate focused on how to ensure against the overissue of notes, so that convertibility could be maintained and commercial crises avoided. The debate was often described both in historical and modern discussions as one in which the Currency School favoured "rules" and the Banking School favoured "discretion." We review the positions of the two sides and argue that because both sides agreed on the passive rule of a gold standard, their dispute is more appropriately characterized as juxtaposing activist and passive rules, rather than rules versus pure discretion.

The Opposing Positions

The Currency School believed that, when paper money and gold were readily convertible, banks frequently issued notes in amounts greater than those under a pure metallic standard. Such "overissues" of notes raised prices and fostered gold outflows, culminating in severe commercial crises. What was required, they argued, was convertibility plus special restrictions on the issuance of bank notes so that a mixed currency of notes and gold fluctuated in amount exactly as a wholly metallic system would have done under identical circumstances—a view called the "currency principle" (Humphrey 1974, p. 7; O'Brien 1992, p. 564).[8]

Members of the Banking School argued that—if bank notes had been issued against the discount of short-term commercial bills drawn to finance real goods in the process of production and distribution—it was not possible for the quantity of money to be excessive, and to thus to cause inflation. Therefore, the nominal quantity of bank notes was determined by the real volume of goods under production, which is why this view became known as the "real bills doctrine" (Mints 1945). In this view, the Bank of England could not force an excess issue of notes on the market, because no one would borrow and pay interest unnecessarily. Any excess would be extinguished as borrowers paid back costly interest-bearing loans to the Bank—an idea known as the "law of reflux" (Humphrey 1988, p. 5). Consequently, the quantity of notes in circulation was adequately controlled by competitive processes. Under convertibility, the quantity of notes would not exceed the needs of business for any appreciable length of time—the "banking principle" (Viner

premium of bullion and the depreciation of the pound sterling following the suspension of convertibility? For discussions, see Humphrey (1974) and Laidler (1992).

[8] Meltzer (2002, p. 36) pointed out that Ricardo had earlier made this argument in his writings and Parliamentary testimony.

1937, p. 223). Thus, members of the Banking School argued that statutory control on the issuance of bank notes was unnecessary.

The debate between the Currency School and the Banking School culminated with the Bank Charter Act of 1844—sometimes called Peel's Act, after then–Prime Minister Sir Robert Peel. The Act marked a triumph for Currency School ideas. The charter split the Bank of England into two departments: an Issue Department and a Banking Department. The Issue Department was limited to an issuance unbacked by bullion—the fiduciary issue—of 14 million pounds (Viner 1937, p. 220). The amount was set considerably below the actual circulation, so that there would be a safe margin backed by gold. Above that amount, the Issue Department could issue notes only in exchange for gold (or, within certain limits, silver).

The Bank of England remained under private ownership and the Banking Department functioned as a private bank. Nonetheless, it occupied a special place in the banking system because the reserves of other London bankers consisted, in part, of deposit balances held on the books of the Banking Department. The Banking Department competed with other banks in providing lending services, but it maintained higher shares of reserves relative to its total liabilities than those banks. Bagehot (1873, pp. 18–19) reported that, in the middle of the nineteenth century, the Banking Department's reserves in bank notes and coin averaged between 30 and 50 percent of its total liabilities, compared with between 11 and 13 percent for other banks.

Under Peel's Act, no new banks could issue banknotes. Existing banks received a compensation if they relinquished the right of issue. Those banks that continued to issue notes were limited to an amount equal to the average circulation in the three months immediately preceding the passage of the Act (Daugherty 1942).

The Rules versus Discretion Terminology

The use of the terms "rules" and "discretion" was commonplace for participants in the Currency School/Banking School debates. For example, the 1840 Parliamentary *Report from the Select Committee on Banks of Issue* heard evidence from ten experts, including Palmer, Norman, Loyd, and Tooke. During the course of the hearings—amounting to some 400 pages—the terms "rule" or "rules" were used 123 times; the term "discretion" was used 18 times.

The widespread use of the terms "rules" and "discretion" in those debates has led a number of modern historians to conclude that the rules-versus-discretion literature originated in the Currency School versus Banking School literature.[9] As one example among many, O'Brien (2007, pp. 98–99) writes:

[9] In the modern literature, similar views to those given in the main text have been expressed by, among others, Laidler (2002, pp. 17–18), Humphrey (1988, p. 4), Flanders (1989, p. 34), Arnon (1991, Chapter 9; 2010), Schwartz (1992, p. 151), and Goodhart and Jensen (2015, p. 21). However, not all contemporary historians identify the Banking School with discretionary policies. As one example, in a paper on the history of rules, Asso and Leeson (2012, p. 8) stated: "[B]oth the Currency School and the Banking School provided cases for subjecting the Bank of England to some preconceived rules of conduct. . . . The Banking School proposed a 'softer' rule."

> Another way of looking at the distinction between the Currency and Banking principles is to view it as a distinction between rules and discretion. . . . [T]he leading member of the Banking School . . . proposed that the Bank of England should hold a gold reserve of between £10 million and £15 million and that it should avoid taking contractionary action on a discretionary basis, only pursuing monetary contraction if the reserve, starting at £15 million, fell below £10 million. . . . The Currency School sought to link the money supply automatically to the balance of payments while the Banking School relied on discretion to avert the catastrophe of a sustained departure from long-run equilibrium values, resulting in the suspension of convertibility.

But a closer look at the how "rules" and "discretion" were used at that earlier time suggests that this modern interpretation is questionable. For example, some Currency School advocates framed their position as favoring rules over discretion. For example, Loyd (1844, p. 21; quoted from Demeulemeester 2019, italics added, p. 80) wrote:

> Without this rule [that is, the currency principle], all must be left to the irregularity and uncertainty of individual discretion. The manager of the circulation must undertake to foresee and to anticipate events, instead of merely making his measures conform to a self-acting test. . . . In the exercise of such a discretion, the manager of the circulation . . . will, in nine cases out of ten, fall into error; whilst the interests of the whole community, and the fate of all mercantile calculations, will be dependent upon the sound or unsound *discretion* of some individual or body; instead of depending upon their own prudence and judgment, exercised under the operation of a fixed and invariable law, the nature and provisions of which are equally known to every body.

Conversely, members of the Banking School sometimes argued that their policy had a discretionary element. Tooke argued that the Bank of England should hold a sufficiently large quantity of reserves so that it could withstand a gold outflow without endangering convertibility. In that way, the Bank would be able to distinguish between a gold outflow that was temporary and self-correcting and an outflow that would be long-lasting, requiring an interest-rate increase. In his *Thoughts on the Separation of the Departments of The Bank of England*, Tooke (1844) set a lower limit for reserves of ten million pounds before the Bank would need to raise interest rates, but he was not always specific about the amount at which the lower limit should be set. In parliamentary testimony in 1848, he was asked about the limit at which the Bank needed to act. He replied (as quoted in Arnon 1991, p. 138): "I am quite sure that you must leave it to the *discretion* of some men or body of men; no doubt they are fallible in their judgement, and Bank directors have sometimes signally failed in their judgement."

In real-world situations, the Currency School and Banking School participants often found themselves in agreement. In the 1847 financial crisis, Bank of

England gold reserves fell from £15 million to £10 million in a period of three months. The Bank responded by raising its discount rate (Daugherty 1943, p. 241), an action consistent with the views of Banking School advocates, including Tooke and Fullarton. However, notice that reserves had not yet fallen below the level of £10 million recommended by the Currency School. Nonetheless, Robbins (1958, p. 119) pointed out that the chief Currency School writers like Torrens and Loyd agreed that a policy of raising the discount rates was "necessary and sensible" in this case because it was a severe emergency. Apparently, on the issue of using the discount rate to respond to exceptional gold drains, the views of the Currency and Banking School essentially coincided.

We believe that the ideas of "rules" and discretion used by participants in the Currency School versus Banking School debate do not correspond to their modern usage. After all, both groups favored a rule—the gold standard rule. In fact, there was no opportunistic, *activist* use of monetary policy preceding or during these debates. Their disagreement was about the best rule for ensuring balance-of-payments adjustment under the gold standard, and in particular, about the use of policy instruments to react to exceptional circumstances—that is, to excessive gold flows. Neither side recognized the possibility that monetary policy could be formulated on the basis of an activist rule that is both systematic and predictable. From a modern perspective, an increase in the discount rate to stem gold outflows once reserves have fallen to a certain level, as under the Banking School framework, does not constitute a "discretionary regime" any more than does a hike in the interest rate in response to a rise in inflation under the Taylor rule.

Intriguingly, the pre-1970s secondary literature on the Currency School versus Banking School debate did *not* interpret that debate within the context of rules versus discretion but, instead, took the position that *both* Schools opposed discretion. For example, Viner (1937, p. 389) wrote:

> Both schools were hostile to discretionary management. The currency school thought that the currency could be made nearly automatic again merely by limiting the issue of bank notes uncovered by specie. The banking school held that there was no acceptable way of escape from the discretionary power of the Bank of England over the volume of deposits, although the "banking principle," according to which the issue of means of payment could not be carried appreciably beyond the needs of business under convertibility, set narrow limits to this discretionary power.

Similarly, Schumpeter (1954, p. 727) argued: "Both [the Currency School and the Banking School] were equally averse to monetary management." Blaug (1962, p. 185) argued: "It is clear that at bottom neither school recognized the necessity of discretionary management of the currency. The Currency School wanted to regulate the note issue . . . while the Banking School balked at the idea of any monetary management whatever." Mints (1945, p. 100) and Robbins (1958, p. 122) expressed similar views.

What happened to produce an about-face amongst recent doctrinal historians compared to the position of their predecessors concerning the applicability of rules versus discretion in the Currency School versus Banking School debate? We conjecture that extensive usage of the terms "rules" and "discretion" by participants in the Currency-Banking School debate has misled some modern historians into taking these terms at face value, rather than recognizing that an activist but well-defined rule is not synonymous with a policy discretion.

Conclusion

This paper has provided a discussion of the evolution (in reverse timeline) of the rules versus discretion debate in monetary policy. All three of the major debates discussed here were initiated in periods of macroeconomic malfeasance and significant monetary policy failures. All leaned strongly, or were decided in favor of, rules. All emphasize the crucial role played by the successful management of expectations for the superiority of rules. And all were cognizant of the fact that a high degree of activism in a rule may create room for discretion—and may even prove counterproductive if it carries excessive informational requirements for the policymakers. In contrast to the two earlier debates, the modern literature seems to have faith in the performance of relatively complex and activist rules—perhaps reflecting the confidence of macroeconomists about their understanding of the economy and their ability to manage the business cycle.

■ *We are especially grateful to Samuel Demeulemeester, Tom Humphrey, Erik Hurst, Ed Nelson, Timothy Taylor, and Heidi Williams for comments. We thank Elisavet Bosdelekidou and Maria Monopoli for excellent technical assistance.*

References

Arnon, Arie. 1991. *Thomas Tooke: Pioneer of Monetary Theory*. Ann Arbor: University of Michigan Press.

Arnon, Arie. 2010. *Monetary Theory and Policy from Hume and Smith to Wicksell: Money, Credit, and the Economy*. New York: Cambridge University Press.

Asso, Francesco Pier, and Robert Leeson. 2012. "Monetary Policy Rules: from Adam Smith to John Taylor." In *The Taylor Rule and the Transformation of Monetary Policy*, edited by Evan F. Koenig, Robert Leeson, and George A. Kahn, 3–62. Stanford: Hoover Institution Press.

Bagehot, Walter. 1873. *Lombard Street: A Description of the Money Market*. Homewood, Illinois: Richard D. Irwin, Inc., 1962.

Barro, Robert J., and David B. Gordon. 1983. "A Positive Theory of Monetary Policy in a Natural Rate Model." *Journal of Political Economy* 91 (4): 589–610.

Blaug, Mark. 1962. *Economic Theory in Retrospect.* Cambridge: Cambridge University Press.
Calvo, Guillermo A. 1978. "On the Time Consistency of Optimal Policy in a Monetary Economy." *Econometrica* 46 (6): 1411–28.
Clerc, Laurent, Harris Dellas, and Olivier Loisel. 2011. "To Be or Not to Be in Monetary Union." *Journal of International Economics* 83 (2): 154–67.
Daugherty, Marion R. 1942. "The Currency-Banking Controversy: Part 1." *Southern Economic Journal* 9 (2): 140–55.
Daugherty, Marion R. 1943. "The Currency-Banking Controversy: Part 2." *Southern Economic Journal* 9 (3): 241–50.
Dellas, Harris, and George S. Tavlas. 2021. "The Dog That Didn't Bark: The Curious Case of Lloyd Mints, Milton Friedman, and the Emergence of Monetarism." *History of Political Economy* 53 (4): 633–72.
Demeulemeester, Samuel. 2019. "The 100% Money Proposal of the 1930s: Conceptual Clarification and Theoretical Analysis." PhD diss. Université de Lyon.
Flanders, M. June. 1989. *International Monetary Economics, 1870–1960: Between the Classical and the New Classical.* New York: Cambridge University Press.
Friedman, Milton. 1948. "A Monetary and Fiscal Framework for Economic Stability." Reprinted in Milton Friedman, ed., *Essays in Positive Economics.* Chicago: University of Chicago Press, 1953, pp. 133–56.
Friedman, Milton. 1953. "The Effects of a Full-Employment Policy on Economic Stability: A Formal Analysis." In *Essays in Positive Economics*, 117–32. Chicago: University of Chicago Press.
Friedman, Milton. 1960. *A Program for Monetary Stability.* New York: Fordham University Press.
Friedman, Milton, and Anna Jacobson Schwartz. 1963. *A Monetary History of the United States, 1867–1960.* Princeton, NJ: Princeton University Press.
Galí, Jordi. 2008. *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework and Its Applications.* Princeton, NJ: Princeton University Press.
Goodhart, Charles A.E., and Meinhard A. Jensen. 2015. "Currency School versus Banking School: An Ongoing Confrontation." *Economic Thought* 4 (2): 20–31.
Humphrey, Thomas M. 1974. "The Quantity Theory of Money: Its Historical Revolution and Role in Policy Debates." Federal Reserve Bank of Richmond, *Economic Review* (3): 2–19.
Humphrey, Thomas M. 1988. "Rival Notions of Money." Federal Reserve Bank of Richmond, *Economic Review* (5): 3–9.
Kydland, Finn E., and Edward C. Prescott. 1977. "Rules Rather than Discretion: The Inconsistency of Optical Plans." *Journal of Political Economy* 85 (3): 473–92.
Laidler, David. 1992. "Bullionist Controversy." In *The New Palgrave Dictionary of Money and Finance*, Vol. 1, edited by Peter Newman, Murray Milgate, and John Eatwell, 255–61. London: Macmillan Press.
Laidler, David. 2002. "Rules, Discretion, and Financial Crises in Classical and Neoclassical Economics." *Economic Issues* 7 (2): 11–34.
Loyd, Samuel Jones. 1837. *Reflections Suggested by a Perusal of Mr. J. Horsley Palmer's Pamphlet: On the Causes and Consequences of the Pressure on the Money Market.* London: Pelham Richardson.
Loyd, Samuel Jones. 1844. *Thoughts on the Separation of the Departments of the Bank of England.* Cornhill: Pelham Richardson.
Meltzer, Allan H. 2002. *A History of the Federal Reserve.* Chicago: University of Chicago Press.
Meltzer, Allan H. 2010. *A History of the Federal Reserve, Volume 2, Book 1, 1951–1969.* Chicago: University of Chicago Press.
Mints, Lloyd W. 1945. *A History of Banking Theory.* Chicago: University of Chicago Press.
Mints, Lloyd W. 1946. "Monetary Policy." *Review of Economic Statistics* 28 (2): 60–9.
Mints, Lloyd W. 1950. *Monetary Policy for a Competitive Society.* New York: McGraw-Hill.
Nelson, Edward. 2020. *Milton Friedman and Economic Debate in the United States, 1932–1972.* Chicago: University of Chicago Press.
O'Brien, Denis P. 1992. "Currency Principle." In *The New Palgrave Dictionary of Money and Finance*, Vol. 1, edited by Peter Newman, Murray Milgate, and John Eatwell, 564–65. London: Macmillan Press.
O'Brien, Denis P. 2007. *The Development of Monetary Economics: A Modern Perspective on Monetary Controversies.* Cheltenham: Edward Elgar.
Parliamentary Committee on Banks of Issue. 1840. *Report from the Select Committee on Banks of Issue.* Great Britain: House of Commons.
Phillips, Ronnie J. 1995. *The Chicago Plan and New Deal Banking Reform.* New York: M.E. Sharpe.
Robbins, Lionel. 1958. *Robert Torrens and the Evolution of Classical Economics.* New York: St. Martin's Press.
Schumpeter, Joseph A. 1954. *History of Economic Analysis.* New York: Oxford University Press.

Schwartz, Anna J. 1992. "Banking School, Currency School, Free Banking School." In *The New Palgrave Dictionary of Money and Finance*, Vol. 1, edited by Peter Newman, Murray Milgate, and John Eatwell, 148–51. London: Macmillan Press.

Simons, Henry C. 1933. "Banking and Currency Reform," including Appendix, "Banking and Business Cycles," and Supplementary Memorandum, "Long-term Objectives of Monetary Management." Unsigned mimeograph, Department of Economics, University of Chicago.

Simons, Henry C. 1936. "Rules versus Authorities in Monetary Policy." In *Economic Policy for a Free Society*, edited by Henry C. Simons, 160–83 and 325–35. Chicago: University of Chicago Press, 1948.

Tavlas, George S. 2022. "'The Initiated': Aaron Director and the Chicago Monetary Tradition." *Journal of the History of Economic Thought* 44 (1): 1–23.

Tavlas, George S. Forthcoming. *The Monetarists: The Making of the Chicago Monetary Tradition, 1927–1960.* Chicago: University of Chicago Press.

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

Taylor, John B. 2017. "Rules Versus Discretion: Assessing the Debate Over the Conduct of Monetary Policy." NBER Working Paper 24149.

Tooke, Thomas. 1844. *An Inquiry into the Currency Principle: The Connection of the Currency with Prices, and the Expediency of a Separation of Issue from Banking.* London: Longman, Brown, Green, and Longmans.

Viner, Jacob. 1937. *Studies in the Theory of International Trade.* New York: Harper and Brothers Publishers.

Walsh, Carl E. 2010. *Monetary Theory and Policy.* 3rd ed. Cambridge, MA: The MIT Press.

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

Journal of Economic Perspectives—Volume 36, Number 3—Summer 2022—Pages 261–268

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by e-mail at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., Saint Paul, MN 55105.

Symposia and E-books

The Aspen Economic Strategy Group has published an e-book with eight essays on the broad theme: *Rebuilding the Post-Pandemic Economy* (November 2021, https://www.economicstrategygroup.org/publication/rebuilding/). As one example, Benjamin F. Jones writes: "We massively underinvest in science and innovation, with implications for our standards of living, health, national competitiveness, and capacity to respond to crisis. . . . Whether facing a pandemic, climate change, cybersecurity threats, outright conflict, or other challenges, a robust capacity to innovate—and to do so quickly—appears central to national security and national resilience. . . . [A] sustained doubling of all forms of R&D expenditure in the U.S. economy could raise U.S. productivity and real per-capita income growth rates by an additional 0.5 percentage points per year over a long time horizon. . . . In many

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.3.261.

ways, the vision of science and innovation needs to be the opposite of 'picking winners.' Rather, we need to 'pick portfolios,' with an emphasis on both increasing the scale of funding and human capital, and the diversity of approaches that are taken." Jones offers many concrete illustrations of broader points: "To study DNA, it must first be replicated into measurable quantities, and this replication process depends on many prior scientific advances. One critical if unexpected advance occurred in 1969, when two University of Indiana biologists, Thomas Brock and Hudson Freeze, were exploring hot springs in Yellowstone National Park. Brock and Freeze were asking a simple question: can life exist in such hot environments? They discovered a bacterium that not only survived but thrived—a so-called extremophile organism—which they named *Thermus aquaticus*. . . . [T]his type of scientific inquiry was motivated by a desire for a deeper understanding of nature, and it had no obvious or immediate application. However, in the 1980s, Kary Mullis at the Cetus Corporation was searching for an enzyme that could efficiently replicate human DNA. Such replication faces a deep challenge: it needs to be conducted at high heat, where the DNA unwinds and can be copied, but at high heat replication enzymes do not hold together. Mullis, in a Eureka moment, recalled the story of *Thermus aquaticus*, knowing that this little bacterium must be able to replicate its DNA at high heat given its environment. And indeed, *Thermus aquaticus* turned out to provide what was needed. Its replication enzyme was declared by *Science Magazine* to be the 'molecule of the year' in 1989. Mullis would be awarded a Nobel Prize soon after, and the biotechnology industry would boom, opening new chapters of human progress."

Recession Remedies: Lessons Learned from the U.S. Economic Policy Response to COVID-19, offers nine essays edited by Wendy Edelberg, Louise Sheiner, and David Wessel (Brookings Institution, April 2022, https://www.brookings.edu/essay/recession-remedies). For example, Edelberg, Jason Furman, and Timothy F. Geithner note: "Overall, the United States' fiscal response appears to have been much larger than the response undertaken by any other country; this was especially true in 2021, when fiscal policy was as supportive as it was in 2020. The U.S. GDP recovery has been among the strongest of any of the advanced economies, but the U.S. employment recovery has been among the weakest; this suggests that both the size of the response and, perhaps, its character and preexisting institutions all matter. . . . The economy experienced major side effects from the pandemic and associated policy response, most notably the highest inflation rate in 40 years, far outpacing the increase in wages and leading to the largest real wage declines in decades. In addition, the U.S. government incurred substantial debt during the pandemic. With the expiration of most forms of fiscal support, real household income is likely to be lower in 2022 than in 2021 and could well be below its pre-pandemic trend. As a result, poverty is on track to rise in 2022. Moreover, inflationary pressures and the efforts to moderate those pressures might bring an end to the expansion." Sheiner's essay looks at state and local spending: "[F]ederal aid was more than sufficient to offset any revenue losses in every state. Nevertheless, state and local government employment declined sharply, and the decline has been quite persistent. . . . [I]n February 2022, the state and local sector accounted for 23 percent

of the shortfall in U.S. employment from its pre-pandemic trend. . . . [G]enerous federal aid to states was clearly not sufficient to reverse or prevent all the employment losses. One important question is, why not? What did state and local governments do with the federal aid, and why didn't they use it to increase employment?"

Justin Sandefur has edited an e-book of six essays on the theme *Schooling for All: Feasible Strategies to Achieve Universal Education*. (Center for Global Development, April 2022, https://www.cgdev.org/publication/schooling-all-feasible-strategies-achieve-universal-education). As one example, Biniam Bedasso writes: "School feeding programs have emerged as one of the most common social policy interventions in a wide range of developing countries over the past few decades. Before the disruptions caused by the COVID-19 pandemic, nearly half the world's schoolchildren, about 388 million, received a meal at school every day (WFP 2020). As such, school feeding is regarded as the most ubiquitous instrument of social protection in the world employed by developing and developed countries alike. But school feeding is also a human capital development tool. . . . A review of 11 experimental and quasi-experimental studies from low- and middle-income countries reveals that school feeding contributes to better learning outcomes at the same time as it keeps vulnerable children in school and improves gender equity in education. Although school feeding might appear cost-ineffective compared with specialized education or social protection interventions, the economies of scope it generates are likely to make it a worthwhile investment particularly in food-insecure areas." In Jack Rossiter's essay, he estimates that the costs of universal primary and secondary school spending would be about $1.9 trillion for low- and middle-income countries in 2030, while the projected education spending for these countries is about $750 billion less. He makes a sobering case: "Even if international financing comes in line to meet targets, governments are not going to have anything like the sums that costing exercises require. We can choose to ignore this shortfall, stick with plans, and watch costs creep up. Or we can see it as a serious budget constraint, redirect our attention toward finding ways to push costs down, and try hard to get close to universal access in the next decade."

The *Journal of Economic Methodology* (2022, 29:1, https://www.tandfonline.com/toc/rjec20/29/1) has published a six-paper symposium for the 50th anniversary of the classic 1972 paper in the *Journal of Economic Theory* by Robert E. Lucas, "Expectations and the neutrality of money." Personal essays by Thomas J. Sargent on "Learning from Lucas" and Harald Uhlig on "The lasting influence of Robert E. Lucas on Chicago economics" describe how Lucas influenced the intellectual journey of the authors. In his essay, Peter Galbács describes how the 1972 paper emerged from Lucas's earlier work: "The way Lucas arrived at his monetary island-model framework was thus a step-by-step process starting in the earliest stage of his career. The first step was the choice-theoretic analysis of firm behaviour. At this stage, Lucas's focus was on the firm's investment decision through which he distinguished short-run and long-run reactions of the firm and the industry. . . . [This work was] shortly extended to labour market modelling—so Lucas's work with [Leonard] Rapping is rooted in his earlier record in firm microeconomics. As they assumed, the household decides on short-run labour supply on the basis of a given

set of price and wage expectations, while it adjusts to long-run changes with a firm-like investment decision that implies the revision of expectations. After this second step taken in labour market modelling, the third stage realizing his Expectations and the neutrality of money (Lucas, 1970/1972a) directly followed. . . . First of all, he needed the very island-model framework. It is [Edmund] Phelps (1970, pp. 6–9) who called his attention to the option of reformulating the decision problem by scattering the agents over isolated markets, while it is [David] Cass who led Lucas to a correct mathematical exposition. However, it is [Edward] Prescott who in their collaboration prepared Lucas for this exposition; and it is also Prescott who, teamed up with Lucas, provided the paradigmatic example of applying the Muthian rational expectations hypothesis in a stochastic setting with which Lucas (1966/1981b) had formerly dealt only in the less interesting non-stochastic case."

Interviews

William Zhao interviews Jeffrey Wooldridge on "The Current and Future of Econometrics" (SciEcon AMA, March 7, 2022, https://medium.com/sciecon-ama/the-current-and-future-of-econometrics-ed30569e7edd, podcast and transcript available). "When we publish papers, the best way to get your work published is to show that it works better than existing methods. Since the people writing the theory and deriving the methods are the same ones doing the simulations, it will probably be better if there's some disconnection there. . . . I've always thought that we should have more competitions, such as blind competitions where people who participate don't know what the truth is. They apply their favorite method across a bunch of different scenarios, so we can evaluate how the different methods do. I'm guessing that machine learning will come out pretty well with that, but that's an impression. I'm not convinced that somebody using basic methods who has good intuition and is creative can't do as well. . . . I think the work on applying machine learning methods to causal inference has guaranteed that it will have a long history in econometrics and other fields that use data analysis. When I took visits to campuses, Amazon, Google, they're using machine learning methods quite a bit. That's no secret. These companies are in the business of earning profits, and they're not going to employ methods that somehow aren't working for them. So, I think the market is certainly speaking on that. For prediction purposes, they seem to work very well."

Noah Smith serves as interlocutor in "Interview: Arvind Subramanian, former Chief Economic Advisor to the Government of India" (Noahpinion, March 31, 2022, https://noahpinion.substack.com/p/interview-arvind-subramanian-former). "Great strides have not just been made in physical but digital infrastructure. In 2015, I coined a term JAM which represented the coming together of financial inclusion (the J from the Hindi "Jan Dhan"), biometric identity (the A for "Aadhaar") and telecommunications (the M for mobile). The government has used this trinity for a variety of purposes, including making direct cash transfers to the poor. In addition, a public-private partnership has created a digital, non-proprietary platform called the Unified Payment Interface (UPI) which is driving a lot of private dynamism

and innovation in a number of sectors—finance, tourism, e-commerce, software solutions etc. I like to joke that India is creating unicorns roughly at the rate that it is creating chess grandmasters. While cause for cheer, this dynamism, based on skill- and technology-intensive factors of production, cannot drive structural transformation because that requires creating jobs for India's vast, relatively less skilled labor force. And India's job situation, especially after the pandemic, is sobering. Which leads to your question about why India has not really managed to achieve scale in its manufacturing operations and why Indian capital is reticent in doing so. I suspect, although I am not sure, that there is again a lot of path-dependence here. For a long time under the license Raj, domestic entrepreneurship was penalized. And there was a particular aversion to size, fearing the economic and political power that large firms could wield. . . . So, paradoxically, labour feels vulnerable to the power exercised by large firms but equally capital does not feel protected by the state either. So, we are in a bad equilibrium that favors small over big."

Allison Schrager has a 50-minute interview with economic historian Joel Mokyr on the topic of "The Future Economy" (Risk Talking podcast, May 17, 2022, https://www.city-journal.org/the-future-economy-with-joel-mokyr). "[T]he real problem is that most of the important contributions to economic welfare are often seriously, seriously, seriously underestimated in our procedures. And I believe that they are getting more and more underestimated. If the degree of underestimation is more or less constant, then you don't care because over time if it isn't changing over time, you can still see what the trend looks like. But I think that's not right. I think we are more and more underestimated because the knowledge economy and the digital economy are famously subject to underestimation. . . . I mean, just look at the enormous gain in human welfare that we have achieved because we were able to come up with vaccines against corona. Now, it's not a net addition to GDP because before that we didn't have corona, but think about the subtraction we would've had if it wasn't for that. And so, I remain a technological optimist, but I'm also very much aware that measures that measure technological progress in a system that was designed for an economy that produced wheat and steel aren't appropriate for an economy that produces high-tech things that are produced by a knowledge economy." Later, Mokyr says: "'Technological progress is neither good nor bad, nor is it neutral.' This is known as Kranzberg's law. It was Melvin Kranzberg who said that, and people keep citing that, although nobody quite knows what he meant."

Sara Frueh interviews Daniel Kahneman in "Try to Design an Approach to Making a Judgment; Don't Just Go Into It Trusting Your Intuition" (*Issues in Science and Technology*, Spring 2022, https://issues.org/daniel-kahneman-interview-noise-judgment-decisionmaking/). "Well, I think that there is widespread antipathy to algorithms, and it's a special case of people's preference for the natural over the artificial. In general we prefer something that is authentic over something that is fabricated, and we prefer something that's human over something that is mechanical. And so we are strongly biased against algorithms. I think that's true for all of us. Other things being equal, we would prefer a diagnosis to be made or a sentence to be passed by a human rather than by an algorithm. That's an emotional thing. But that feeling has to be weighed against the fact that algorithms, when they're

feasible, have major advantages over human judgment—one of them being that they are noise-free. That is, when you present the same problem to an algorithm on two occasions, you are going to get the same answer. So, that's one big advantage of algorithms. The other is that they're improvable. So, if you detect a bias or you detect something that is wrong, you can improve the algorithm much more easily than you can improve human judgment. And the third is that humans are biased and noisy. It's not as if we're talking of humans not being biased. The biases of humans are hidden by the noise in their judgment, whereas when there is a bias in an algorithm, you can see it because there is no noise to hide it. But the idea that only algorithms are biased is ridiculous; to the extent they have their biases, they learn them from people. . . . The real deep principle of what we call decision hygiene is independence. That is, you want items of information to be as independent of each other as possible. For example, you want witnesses who don't talk to each other, and preferably who saw the same event from different perspectives. You do not want all your information to be redundant. So, good decisions are decisions that are made on the basis of diverse information." Kahneman also notes: "I have more confidence in the ability of institutions to improve their thinking than in the ability of individuals to improve their thinking."

Discussion Starters

Angelo Duarte, Jon Frost, Leonardo Gambacorta, Priscilla Koo Wilkens, and Hyun Song Shin tell the story of "Central banks, the monetary system and public payment infrastructures: lessons from Brazil's Pix" (Bank of International Settlements, BIS Bulletin #52, March 23, 2022, https://www.bis.org/publ/bisbull52.pdf). "The BCB [Brazil Central Bank] decided in 2018 to launch an instant payment scheme developed, managed, operated and owned by the central bank. Pix was launched in November 2020. . . . The BCB plays two roles in Pix: it operates the system and it sets the overall rulebook. As a system operator, the BCB fully developed the infrastructure and operates the platform as a public good. As rulebook owner, the BCB sets the rules and technical specifications (e.g., APIs) in line with its legal mandate for retail payments. This promotes a standardised, competitive, inclusive, safe and open environment, improving the overall payment experience for end-users. Since its launch, Pix has seen remarkable growth. By end-February 2022 (15 months after launch), 114 million individuals, or 67 percent of the Brazilian adult population, had either made or received a Pix transaction. Moreover, 9.1 million companies have signed up—fully 60 percent of firms with a relationship in the national financial system."

Manvir Singh describes the evidence in "Primitive communism: Marx's idea that societies were naturally egalitarian and communal before farming is widely influential and quite wrong" (*Aeon*, April 19, 2022, https://aeon.co/essays/the-idea-of-primitive-communism-is-as-seductive-as-it-is-wrong). "The idea goes like this. Once upon a time, private property was unknown. Food went to those in need. Everyone was cared for. Then agriculture arose and, with it, ownership

over land, labour and wild resources. The organic community splintered under the weight of competition. The story predates Marx and Engels. The patron saint of capitalism, Adam Smith, proposed something similar, as did the 19th-century American anthropologist Lewis Henry Morgan. Even ancient Buddhist texts described a pre-state society free of property. . . . Today, many writers and academics still treat primitive communism as a historical fact. . . . Primitive communism is appealing. It endorses an Edenic image of humanity, one in which modernity has corrupted our natural goodness." After a review of property rights, punishments, and some examples of brutal behavior in early societies, Singh writes: "For anyone hoping to critique existing institutions, primitive communism conveniently casts modern society as a perversion of a more prosocial human nature. Yet this storytelling is counterproductive. By drawing a contrast between an angelic past and our greedy present, primitive communism blinds us to the true determinants of trust, freedom and equity."

Nuno Palma, Andrea Papadia, Thales Pereira, and Leonardo Weller discuss "Slavery and Development in Nineteenth Century Brazil" (*Capitalism: A Journal of History and Economics*, Summer 2021, 2:2, pp. 372–426, https://muse.jhu.edu/article/798739). "Prior to abolition in 1888, slavery was a pronounced and pervasive feature of Brazil's economy. More African captives arrived on Brazilian shores than anywhere else in the Americas. From the sixteenth to the nineteenth century, 4.9 million Africans landed in what was a Portuguese colony in the Americas until 1808, an independent joint kingdom with Portugal from 1808 to 1822, and then the Brazilian Empire from 1822 until the Republic was proclaimed in 1889, the year after emancipation. The total number of Africans transported to Brazil corresponds to 46 percent of all the enslaved arrivals in the New World and double the number who arrived in the whole of the British Caribbean. In comparison, the slave trade to the United States was much smaller: only 388,746 slaves disembarked there . . . [T]he abolition of slavery allowed municipalities to exploit their potential to become manufacturing centers. . . . This result also highlights the presence of potential distortions in the Brazilian economy brought about by slavery: locations with high potential for industrialization, as evidenced by post-abolition developments, were actually disadvantaged earlier on due to a continued focus on cash crops fueled by the prevalence of slave-based production. If we consider the fact that slavery discouraged free migrants from settling, slavery might have also been harmful through this additional indirect channel. . . . There is no evidence that slavery benefited the societies that relied largely on it. Not only is slavery abhorrent from a modern normative perspective, but it also mostly had negative development consequences: while slave-owners and a few narrow sectors profited from it, overall society lost out. . . . The case of Brazil lends credibility to the view that slavery benefited a small elite but delayed overall economic development in the societies where it existed, as has been argued for the US South."

Jeffrey Brinkman and Jeffrey Lin discuss "The Costs and Benefits of Fixing Downtown Freeways," subtitled "Urban freeways spurred our suburban boom. Can burying them do the same for the urban core?" (*Economic Insights: Federal Reserve*

Bank of Philadelphia, Winter 2022, 7:1, pp. 17–22, https://www.philadelphiafed.org/the-economy/the-costs-and-benefits-of-fixing-downtown-freeways). "Using fine geographic data covering 1950 to 2010, we studied long-run changes in neighborhoods before and after the interstate highway system was built. . . . We find that in the group of central-city neighborhoods closest to freeways, population declined by 32 percent, while in the group of central neighborhoods more than 2 miles from freeways, population actually grew by 56 percent." "Using quantitative methods developed in urban economics, we simulate the effects of burying a section of I-95 from Snyder Avenue to Girard Avenue [in Philadelphia]. This roughly 4.5-mile stretch of freeway starts in South Philly and traverses the riverfront neighborhoods of Pennsport, Queen Village, Society Hill, Old City, Northern Liberties, and Fishtown. The proximity of these neighborhoods to the central business district and their high population density suggest that this might be an ideal setting for such an intervention. . . . Economic development was an important rationale for freeway construction, but not everyone benefited from the new freeways. That's because freeways bring amenities to some neighborhoods by increasing access but disamenities to others by reducing the quality of life. Using techniques developed in recent economic research, we can quantify neighborhood amenities and thus the costs and benefits of freeway construction for individual neighborhoods and for an entire metro area. Many cities, including Philadelphia, could benefit from mitigation of freeway disamenities by covering or capping central city highways."

Áine Doris asks "Do Shoppers Have Too Many Choices? US consumer goods are proliferating rapidly, with implications for consumers and companies" (*Chicago Booth Review*, May 23, 2022, https://www.chicagobooth.edu/review/do-shoppers-have-too-many-choices). "The number of 'niche' alternative products increased by 4.5 percent a year from 2004 to 2016, according to a study of consumer packaged goods [Joseph] Vavra conducted with Booth's Brent Neiman in which they analyzed data on almost 700 million transactions involving 118 different product groups. US households appear to have welcomed this product explosion. . . . Growing variety, when all else is equal, creates what economists call positive welfare effects for consumers. As companies fragment their products more and more, consumers are able to buy the things that they really like—getting closer to their optimal choice. Even better for consumers, we haven't been paying more for the additional choice. . . . Over time, the likes of General Mills, Nestlé, Procter & Gamble, and Unilever have systematically acquired and subsumed other consumer brands. This trend notionally gives them not only market share but also the lion's share of market power by which to influence or even set prices. . . . [But] this trend is being offset by another: an upswing in competition at the individual product level. Although there are fewer companies offering products in a certain sector (say, food products), there are more companies offering them in a specific market (such as chips). That is creating more competition at the level of individual products, which keeps prices low."

The *Journal of Economic Perspectives*: Proposal Guidelines

Considerations for Those Proposing Topics and Papers for *JEP*

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.

Philosophy and Style

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of *JEP* articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author.** The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

JEP does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-U.S. country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives.* Our stock in trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship. If you are not familiar with this journal, it is freely available on-line at www.aeaweb.org/journals/jep.

Guidelines for Preparing *JEP* Proposals

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP.* However, given

the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion.
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and authors an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article. After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future articles for *JEP* may be sent (electronically please) to the assistant managing editor, Alexandra Szczupak at a.szczupak@aeapubs.org. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

Guidelines for Empirical Papers Submitted to *JEP*

JEP is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1) The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.

2) In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.

3) The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.

JOE NETWORK

The career network designed by economists—for economists.

The AEA's JOE Network is the preferred hiring tool for the economics job market, matching qualified candidates with relevant economics positions in leading institutions. Over 1,700 positions are filled each year from a pool of 5,100+ candidates.

Job Seekers

Showcase your profile to hundreds of hiring managers and quickly apply to your preferred positions. The JOE Network allows you to securely request reference letters and share job market materials.

Employers

Target search requirements to locate your preferred candidates, set up hiring committee access, and easily manage job postings and applications.

Faculty Letter-Writers

Reply to reference requests with default or customized letters, assign proxies, and easily monitor task completion. Over 150,000 requests are fulfilled on the JOE Network each year.

Use the JOE Network as your comprehensive tool for the economics job market!

AMERICAN ECONOMIC ASSOCIATION

aeaweb.org/JOE

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary access to JEP articles, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $24.00, $34.00, or $44.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

The Journal of

Economic Perspectives

Summer 2022, Volume 36, Number 3

Symposia

Intangible Capital

Human Capital

Inflation Expectations

Methods in Applied Micro

Article

Features

AMERICAN ECONOMIC ASSOCIATION