The Journal of Economic Perspectives

A journal of the American Economic Association

Spring 2025

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 at Sheridan Press, Hanover, Pennsylvania, USA.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Spring 2025, Vol. 39, No. 2. The JEP is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.

POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the USA.

The Journal of
Economic Perspectives

Contents

Volume 39 • Number 2 • Spring 2025

Symposia

Statement of Purpose

The *Journal of Economic Perspectives* aims to bridge the gap between the general interest business and financial press and standard academic journals of economics. 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 39, Number 2—Spring 2025—Pages 3–26

Economic Markets and Pharmaceutical Innovation

Craig Garthwaite

Public opinion polling by Gallup (2024) places the pharmaceutical industry's net favorability rating at negative 42 percent. This ranks below the legal field, oil and gas industry, and the federal government. However, by other measures, the pharmaceutical industry should rightly be viewed as an enormous success story. After all, prescription drugs occupy a central place in treating, managing, and curing a host of chronic and acute conditions. Pharmaceutical innovations are responsible for 35 percent of the remarkable decline in cardiovascular mortality from 1990 to 2015 (Buxbaum et al. 2020). Previously deadly conditions such as HIV/AIDS have been transformed into manageable chronic maladies and others such as hepatitis-C have been cured. Gene therapies are becoming more commonplace as treatments for a wide range of rare and deadly genetic conditions. Advancements in immuno-oncology are providing meaningful advances across a variety of cancers as the body's natural systems are used to combat cancer.

Most recently, the first truly effective treatments for obesity in the form of GLP-1 agonists have emerged with corresponding improvements across a host of cardiometabolic outcomes such as heart disease, diabetes, and chronic kidney disease. In many ways, these new obesity treatments reflect the fundamental

■ *Craig Garthwaite is Professor of Strategy, Kellogg School of Management, Northwestern University, Evanston, Illinois. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is c-garthwaite@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.20251438. The author has received financial support from an interested party.

tension at the heart of pharmaceutical markets. Few dispute their benefits. But many reasonably ask: At what price? This dichotomy is likely true across all product categories.

The high prices for brand-name drugs are no illusion—many recent innovations are sold at eye-watering prices. However, such prices are not an accident, nor should they be unexpected. Instead, they result from policies intended to provide financial incentives for the continued development of new innovative products. Such policies are necessary because of the market failure central to pharmaceutical innovation. Firms developing novel products create public goods in the form of the scientific knowledge necessary to treat a medical condition. After such knowledge is developed, it is often a relatively trivial matter for another firm to replicate the product at a meaningfully lower fixed cost and drive down the financial return for the original innovator.

Therefore, absent some form of intellectual property protection, few rational firms would be willing to make the investments essential to product innovation. This is particularly true in markets, such as pharmaceuticals, where the costs of development are quite large, risks of failure are meaningful, intended payoff periods are long, and the costs of imitation are low. To address this concern, the current system involves governments providing intellectual property protection for limited time periods, so that innovative firms can sell their products at a higher price without the threat of direct competition.

When prices are high, intellectual property protections reduce access to existing innovations. However, the high prices also provide financial incentives for future investments in developing new products. While these short-run and long-run tradeoffs of patents hold true for all innovative markets, they somehow seem more salient in pharmaceuticals. Witness that the high prices and profit margins resulting from intellectual property protection for other innovative products such as semiconductors receive relatively little attention in the popular discourse despite having industry dynamics quite like pharmaceuticals.

The tradeoffs central to drug development quite understandably cause stronger feelings from multiple perspectives and stakeholders. In the short run, a phrase such as "reduced access" is an impressively sterile way of describing the potentially devastating health consequences for readily identifiable patients today that might not access a new drug treatment because of its high price. Of course, those readily identifiable patients are not the only market participants that may have difficulty treating their medical conditions. Patients suffering from conditions without any existing treatments lack access at any price—a problem that might best be solved by the incentives resulting from high prices for existing products

This essay provides a broad summary of our existing knowledge about several important economic facets of the pharmaceutical market as they relate to the innovative process. The hope is that those interested in or directly engaged in debates over the pharmaceutical industry can benefit from understanding fundamental economic facts and patterns surrounding the incentives for drug development.

Market Size and Drug Development Incentives

Developing innovative pharmaceutical products requires large, fixed, and sunk investments in research and development (R&D) activities. The funds for these investments come from private firms, universities, and public institutions. Many successful products often start their path to market in academic labs and small biotechnology firms that hope to license or sell their innovation to larger firms. Even universities that are often thought of as pursuing "pure" or "basic" science have developed extensive transfer and licensing operations. Many of these scientists conducting this early-stage research ultimately form private firms aimed at future commercial development. In all of these ways, the potential commercial success for a new drug is intrinsically tied to the frequency and magnitude of investments in even the earliest stages of development.

Private-sector investments in pharmaceutical research and development are quite large. In 2021, pharmaceutical firms around the globe invested a total of $276 billion, or 27 percent of global revenue, in R&D (Chandra et al. 2024). Unsurprisingly, the largest pharmaceutical industry investments were made by publicly-owned commercial firms, which tend to be larger firms actively selling products. These firms accounted for $244 billion of the total, with $44 billion at the pre-market research stage and $197 billion at the commercial development stage. Privately-held pharmaceutical firms—often hoping to sell or license their efforts to the larger companies for commercialization—accounted for almost $26 billion of "pre-revenue" research spending and only $6 billion of commercial-stage development (Chandra et al. 2024).

Numerous empirical studies show a relationship between a drug's expected market size and the magnitude of research and development investments. Early studies focused on changes to market size resulting from the demographics of disease burden (Acemoglu and Linn 2004) and policy changes influencing market demand (Finkelstein 2004). These findings have largely been confirmed by more recent papers using changes in the generosity of insurance markets as shocks to market size (Blume-Kohut and Sood 2013) or variation in demographics related to medical conditions (Dubois et al. 2015). For example, DuBois et al. (2015) find that each new drug requires an increase in market size of approximately $2.5 billion.

Demonstrating the centrality of financial incentives to research and development investments, a series of studies have documented that firms respond to the potential *economic* size of a product's market and not simply the number of potential patients. For example, Garthwaite, Sachs, and Stern (2022) examine the large Medicaid expansions that occurred as part of the Patient Protection and Affordable Care Act of 2010. Because Medicaid pays much lower prices than other US insurers, this large expansion had only a modest increase in revenue and did not result in increased investments. In another study, Agha, Kim, and Li (2022) exploit the increased use of stronger bargaining tactics in the form of excluding coverage for specific pharmaceutical products. Such tactics did not reduce the number of

patients, but simply the expected financial return per potential patient. The authors found reduced R&D investments in products for therapeutic areas where there are already competing therapeutic substitutes, because new products could easily be pitted against existing drugs during price negotiations.

The centrality of financial incentives to research and development has global implications. The gap between pharmaceutical prices in the United States and other developed markets is well documented and heavily discussed across a variety of policy settings. However, many other developed (and developing) markets are likely too small to be central to the drug development investment decisions of innovative firms. This may explain why many of these developed markets can negotiate access to drugs at prices far lower than in the United States without meaningful fears of causing a reduction in future innovation (Lakdawalla 2018).

Although research and development investments clearly respond to market opportunities, the exact benefits created by these incremental investments remain unclear. This is largely because the exact marginal products developed in response to these incentives are elusive to identify. Certainly, the amount of R&D spending or a simple count of the number of resulting new products is an incomplete metric of social welfare benefits.

One effort to understand the social benefit of incremental products has focused on the scientific novelty of these marginal drugs. Dranove, Garthwaite, and Hermosilla (2022) examine the scientific novelty of the targets of clinical trials launched because of the increase in market size resulting from the creation of Medicare Part D (that is, the social insurer for prescription drugs for the elderly). They find the products involved in these additional trials were not novel in terms of their scientific approach, but they did represent some novel combinations of existing scientific approaches—which could provide new benefits to the market. For example, treatment effects or adverse events may vary across patients among even those products with quite similar scientific approaches (Jena et al. 2009).

In contrast, Krieger, Li, and Papanikolaou (2022) exploit the fact that creating Medicare Part D caused different immediate impacts on the balance sheets of firms with products already on the market. They find that firms enjoying a larger cash shock from Part D increased their investments in products that are more scientifically novel compared to firms not enjoying similarly large cash infusions. Importantly, this effect is not driven by the potential returns of the new products but instead by the cash infusions reducing frictions in the financing market for new innovations.

These two conflicting sets of results suggest those primarily interested in promoting the most scientifically innovative process must concentrate on the specific impediments to such products being developed. For example, if firms are risk-averse due to a costly external financing market, increases in cash reserves or reforms to this external market relax this constraint, causing firms to make riskier investments in more novel products. However, if the concern is something broader, such as a lack of basic science upon which to build commercial product platforms, perhaps more support for such research endeavors should be considered.

The topic of novelty and welfare gains from additional drug research and development is ripe for additional research. One caution that remains is that the welfare effects of new products cannot be fully encapsulated by scientific novelty. While scientifically novel products have higher expected net present values, such novelty is hardly a sufficient statistic for economic welfare. New applications of existing mechanisms of action can have meaningful impacts on patient health and economic welfare.

What Determines Revenue for Drug Innovations

In one way, the potential revenues for a new pharmaceutical product are trivial to understand: it is just the potential quantity sold multiplied by an expected price. However, determining factors such as who exactly the customer is, what specifically the "price" is, and the expected market size are often less clear in drug markets than in more traditional product market settings. Therefore, estimating potential drug market returns becomes an exercise in understanding critical features such as the price-setting mechanism, the determinants of consumer demand, and the institutional factors influencing those components. This section provides an overview of institutional features of the market for new drugs and how they shape revenues.

An Overview of the Value Chain in the Pharmaceutical Market

Every commercially successful pharmaceutical innovation must accomplish three tasks: (1) it must be approved by a regulator; (2) it must be accepted for payment by a third-party payer; and (3) it must be adopted by clinicians (and their patients).[1] These activities are clearly interrelated. For example, the information generated throughout the clinical trial process and the prices negotiated by third-party payers will influence the coverage and clinical adoption decision of payers and providers.

There are many detailed summaries of the intricacies of the market and the various actors involved in the negotiation and product delivery space (for example, Garthwaite and Starc 2023). Here, I will focus on two different specific value chains. The first is for "retail pharmaceuticals," which are prescribed by a doctor and purchased by patients at a pharmacy. The second is for "physician-administered" drugs that are prescribed and given to patients in a medical office.

With retail drugs, perhaps the most important economic feature of the value chain is the relationship between the third-party payers (the actual entity responsible for medical costs), the "pharmacy benefit manager" firms that they hire to handle their drug reimbursements, and drug manufacturers.

[1] These factors are distinct from "health process innovations," which involve new organizational methods that a healthcare organization might use to improve efficiency and outcomes. Such changes do not require the approval of the regulator, which can shape the incentives for innovation in processes compared to products (Dranove et al. 2022).

Every pharmaceutical product enters the retail market with a publicly announced "list" price. The pharmacy benefit manager, as the agent of the third-party payer, negotiates with the manufacturer for a discount, commonly referred to as a "rebate." For each product, rebates vary depending on the relative bargaining power of the manufacturer and the pharmacy benefit manager. Actual rebates are kept strictly confidential in an attempt to facilitate larger negotiated discounts.

Pharmacy benefit managers and drug manufacturers also negotiate the extent of "utilization management," with products from manufacturers providing larger discounts being exposed to fewer of these tools. Pharmacy benefit managers can use nonfinancial tools such as requirements that patients receive "prior authorization" from the third-party payer in order to be reimbursed for a drug, or "step therapy" where a patient must, for example, try a version of a drug preferred by the third-party payer before being eligible for reimbursement for another drug. Utilization management can also involve financial tools like deductibles, coinsurance, and copayments.[2]

Pharmacy benefit managers are compensated under a variety of contracts and therefore receive revenue from numerous streams. At a high level they earn the majority of their income from a combination of payments: (1) a flat fee that is normally calculated per member per month; (2) a percentage of the negotiated rebate; and (3) the spread between what the pharmacy benefit manager pays to the pharmacy for a product and what the plan sponsor pays the pharmacy benefit manager—a practice known as "spread pricing." These features can be controversial. For example, there are sometimes calls for "delinking" compensation of pharmacy benefit managers from rebates and negotiated discounts, perhaps in favor of a system of flat fee compensation for these firms. But of course, these forms of compensation also give pharmacy benefit managers a direct incentive to negotiate for greater discounts (Mulligan 2023).

Historically, the pharmaceutical market was almost entirely "small molecule" products flowing through the value chain described above. In recent years, there has been growth in biologic products that are often administered by physicians and get to patients through a largely different value chain. In 2021, such physician-administered drugs accounted for approximately 27 percent of overall drug spending. This is an increase of nearly 100 percent since 2016 (Garthwaite and Starc 2023). This growth partly reflects the increased development of more complicated medications that require infusion rather than an oral or subcutaneous method of administration.

In contrast to retail drugs, physician-administered drugs are paid for under the "medical" portion of a patient's insurance benefit and therefore do not involve pharmacy benefit managers. For these products, health care providers are responsible

[2] These are the primary forms of cost sharing in pharmaceutical markets. Deductibles are the portion of spending where the patient has sole financial responsibility. Copayments are fixed payments a patient must make in order to obtain a prescription. Coinsurance is cost sharing that is a percentage of the list price of the pharmaceutical.

for acquiring, stocking, and administering the product. When providers prescribe and administer the product to a patient, they are normally paid a mark-up over the product's average sales price, which is the net price paid by providers after manufacturer rebates. In Medicare, physicians are paid 106 percent of the average sales price of the drug. Private market reimbursements typically involve larger markups.

This form of payment can have broad implications. At a minimum, physicians receive greater revenue from prescribing higher-priced products—which can distort new product entry and diffusion. This has certainly been a policy concern for incentives for the development of "biosimilar" products, which are effectively generic forms of competition for biologic products that have lost patent protection. Given these concerns, many payers provide physicians administering biosimilars a payment equal to the original product.

Cost-plus reimbursement for physician-administered drugs also has implications for provider market structure. Depending on the medical specialty, the profits from these drugs can be more or less important to the finances of the medical practice. For example, in the market for dialysis services payments for drugs under Medicare Part B had wide-ranging effects on market structure and firm behavior (Eliason et al. 2020). All else equal, larger health care providers can negotiate bigger discounts. As a result, the ability to secure greater discounts if a practice is part of a larger health system may influence optimal decisions for firms about consolidation.

The value chain of physician-administered drugs can lead to additional market distortions. For example, if a provider deals with multiple third-party payers, then decisions about how or whether to cover prescription drugs by one payer can influence whether physicians find it optimal to stock the drug and by extension make it available to patients covered by other payers. Dean, Feng, and Maini (2024) show that increases in the share of formulary exclusions for a product in the commercial market causes a decrease in the market share in the Medicare Part B market. In addition, implementing a "closed formulary"—in which only drugs on a certain list are eligible for reimbursement—can be more difficult, because patients may face more frictions accessing products if it requires moving across medical providers.

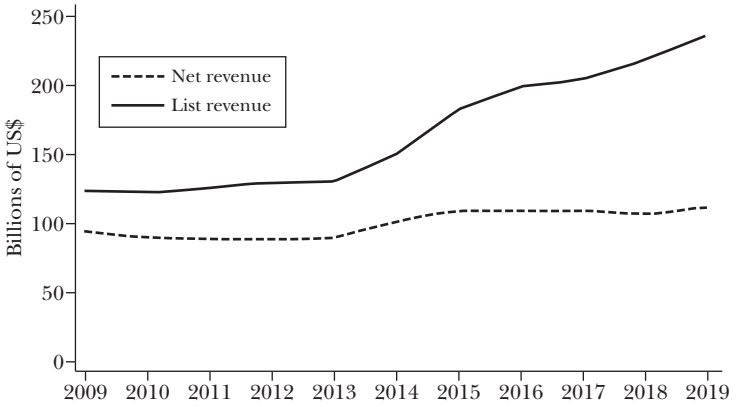
How the Negotiations Determine Price Captured by Drug Manufacturers

While rebates from drug manufacturers have always been a central feature of the price negotiation process, they have grown in magnitude (and therefore strategic importance) in recent years. As shown in Figure 1, the spread between list and net prices for retail pharmaceutical products from 2009 to 2013 was largely stable. But then, list revenue escalated rapidly and the net revenue remained largely constant. By 2019, the magnitude of rebates doubled and the average rebate was approximately 53 percent (Kakani, Chernew, and Chandra 2022).

The list price and the rebate for a particular drug are ultimately the results of the competitive environment in the therapeutic class. Relatively unique therapies often have higher net prices that more closely match list prices. Figure 2 illustrates two of the patterns that can emerge from these negotiations.

Figure 1
The Spread between List and Net Prices for Retail Pharmaceuticals

Panel A. Annual rebate growth



Panel B. List and net price inflation

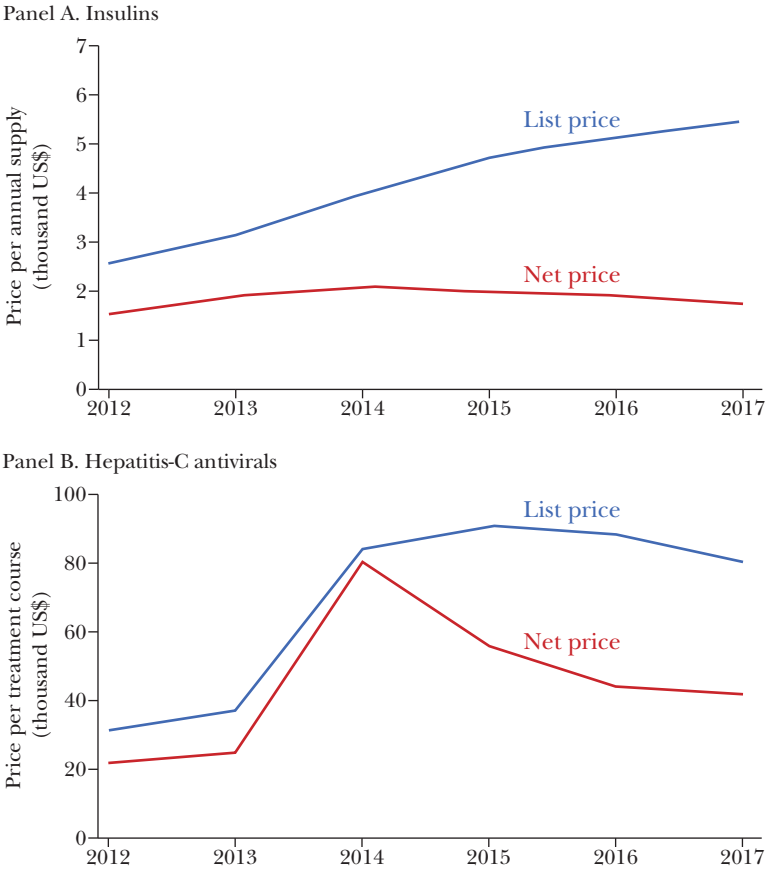
												Average annual growth		
Year	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*09–14*	*14–19*	*09–19*
Net rebates	24%	26.5%	29.3%	31.7%	31.2%	32.6%	40%	44.8%	46.6%	61.1%	52.6%	1.7 pp	4 pp	2.9 pp
Sample (*N*)	841	866	871	774	796	792	778	776	740	684	611			

Source: Kakani, Chernew, and Chandra (2022).
Note: Both panels reflect analysis of 1,374 branded product formulations sold in primarily retail pharmacies, but the sample varies in each year because of entry and exit. See Kakani, Chernew, and Chandra (2022) for additional details on exclusions. Panel B reflects list and net price inflation calculated with a Laspreyres price index for each year pair. This reflects a chain-weighted approach using the balanced sample of products in each pair of adjacent years. Annual inflation is compounded year over year to estimate cumulative inflation relative to 2009. Prices in 2009 are benchmarked at 100 percent.

Figure 2, panel A, contains list and net prices for insulin—a treatment for diabetes. From 2012 to 2017, despite massive increases in list prices, there was effectively no change in net prices. In contrast, Figure 2, panel B, shows prices for treatments for hepatitis C. For these prices, there was a meaningful increase in list and net prices in 2014 followed by a precipitous decline. In 2014, Gilead Sciences released a revolutionary cure for hepatitis C: in that year, this product was the only available cure for the condition. Then in early 2015, the pharmaceutical firm Abbvie released a competitor product with nearly similar efficacy, giving payers the ability to swiftly negotiate larger rebates.

This variation in list and net prices across products demonstrates the importance of therapeutic substitutes for introducing price competition. The benefits of new products can extend well beyond increased competition. While it is often assumed that the second and third drugs to market provide little incremental

Figure 2
List and Net Price: Two Examples



Source: Kakani, Chernew, and Chandra (2022).
Note: Figure 2 depicts estimates of the price for a course of treatment for Insulin (panel A) and for antivirals for Hepatitis C. Each panel contain a price estimate based on the publicly available list price and the "net price," that is, the price after negotiations. Variations in these prices over time are the result of both the introduction of new products and the outcomes of annual negotiations.

clinical benefit, this is not always true. Across a wide variety of medical conditions, the first drugs introduced were important scientific contributions, later supplanted by second- and third-generation products providing superior overall efficacy, fewer side effects, or both. In addition, newer-generation products may prove superior for specific patient groups. In such situations, the result may be that prices do not fall with entry, but each version of drug ends up treating a smaller but more targeted population.

This possibility is not simply theoretical. Jena et al. (2009) examine how demand among substitute products changes when the first product in a class loses

patent protection. Prices for incumbent products plummet following patent expiration. If second- and third-generation products were quite close substitutes for the original, then their prices and/or revenues would also fall. However, across several well-defined disease areas, the authors find no such decreases in price or quantity for the next generation products. They also demonstrate this is not simply a question of inertia of existing patients. Looking only at new prescriptions for patients not previously taking any medication, there is continuing demand for second- and third-generation products.

Specifying the Components of Willingness-to-Pay for Pharmaceuticals

Patent protection is of course intended to provide higher margins for innovators. However, patents are meant to exist within a broader market-based framework where manufacturers must still negotiate prices with payers in competition with other products. But if patents are combined with third-party insurance payers—who shield market participants from transaction prices—then a concern arises that drug manufacturers may gain an inefficiently large amount of power in these price negotiations. In some instances, the existence of health insurance across a variety of products can allow manufacturers to capture more economic value than their specific products create (Besanko, Dranove, and Garthwaite 2020). Understanding the reasons why such value capture occurs is important for understanding welfare implications.

For understanding a product's value, the gain in health received by patients taking the drug is a logical starting point. If products were purchased directly by patients using only their own resources, then this clinical value might even be a sufficient statistic for characterizing that market's willingness-to-pay. A common approach for quantifying clinical value is to calculate how much a product increases quality-adjusted life years or QALYs, which combine any gain in overall life expectancy, along with weighting the improved quality of life on a scale ranging from 0 (death) to 1 (perfect health). Such technology assessments have been used to determine so called "value-based" prices for new products (Yeung et al. 2022)

There is considerable controversy around using quality-adjusted life years as a measure of clinical value. However, when it comes to the impact on pricing, such debates often miss the more important *economic* question about what other components determine the *market's* willingness-to-pay for new innovations. The importance of a broader perspective is particularly important when we consider that pharmaceutical products are not purchased directly by consumers. Instead, consumption is financed by the resources of the entire insurance risk pool, many members of which do not directly benefit from the transaction. In such market settings, there are other potential sources of value for new drugs that are missed by a focus on clinical quality or other features that benefit only the patient and/or their family. Examples include an option value for patients, scientific progress for future treatments, and the insurance value of medical innovation (Lakdawalla 2018).

Consider how medical innovations can change available treatment options for individuals who are not yet afflicted, but could become sick in the future. New

medical technologies transform the medical risk individuals face (that is, becoming afflicted with a condition for which there is no treatment) into a financial risk (that is, finding a way to finance the purchase of medical innovations if they get sick) (Lakdawalla, Malani, and Reif 2017). All risk-averse consumers should value this reduction in health variance. Indeed, the insurance value of the new innovation can even exceed the value of health insurance in the first place, especially for disease areas where the existing treatment armamentarium is quite poor and the physical effects of the condition are quite severe. This could explain why many treatments for rare diseases so often exceed several thresholds based solely on clinical value.

Another gain from new drugs is that scientific progress is often iterative, building on the knowledge and insights from previous advances. Thus, an optimal level of innovation will only be achieved to the extent the eventual value created for society by the next generation of innovations is in some way accounted for in revenues for the manufacturers making incremental progress. Conceptually, this is an area where an efficiently designed government policy could provide such incentives. However, this requires defining clear rules for potential future rewards, which can be quite difficult in practice.

Understanding the potentially broader economic value created by new drug products is not simply an academic exercise. In the short run, "value-based" efforts to link drug prices to clinical outcomes may neglect some benefits of innovation and thus create artificially low prices for existing products (Jena and Philipson 2007). These low prices could alter incentives to invest in developing products in those areas. Concerns over whether drug manufacturers have too much negotiating power over price in a market with third-party payers need to be evaluated in this context.

Commercialization Efforts

Manufacturers attempt to expand the potential revenue for their product by marketing to patients and providers. Such efforts are governed by a variety of Food and Drug Administration regulations. At a minimum, in these marketing efforts, firms are only allowed to discuss medical indications for which the product has been approved. This is despite the fact that physicians are allowed to use approved products for any medical indication and that many novel products have shown efficacy across a variety of conditions. As a result, firms must decide for which indications they should seek approval. Obtaining approval for such indications has been shown to increase the amount of market information about a product and increase use beyond the rate of market learning that otherwise would have occurred (Berger, Chandra, and Garthwaite 2021).

Despite the fact that marketing activities are strictly governed by the Food and Drug Administration, efforts by manufacturers to increase demand remain controversial. One set of concerns revolves around questions of whether such expenditure represent waste and therefore society could receive a similar level of innovation at lower prices (for example, Angelis et al. 2023). However, this view of marketing and R&D expenditures as substitutes misconstrues the underlying economics. If

commercialization increases the potential market size of new innovations, then it encourages firms to invest in developing products and is actually a complement to R&D. If we are willing to assume that firms are correct in believing that commercialization activities, on average, do expand markets, then the key question becomes why consumption was previously lower than the post-commercialization level, because the source of this difference will influence judgements about the value of pharmaceutical marketing efforts.

A broader set of concerns revolves around the more general question of the appropriateness of promoting the increased use of medical services—although it does seem important to note that such concerns are rarely voiced for the advertising of medical services by providers such as hospitals (Starc 2023). Understanding more about the validity of these concerns requires more carefully examining the welfare effects of these commercialization activities.

Marketing Pharmaceutical Products Directly to Customers

Firms in all product markets routinely engage in direct-to-consumer advertising as they attempt to increase sales; however, only the United States and New Zealand allow such advertising in health care markets. A central question around all advertising (Bagwell 2007), but perhaps especially around health care advertising, is whether it primarily represents information or persuasion. In particular, there are concerns that such advertising for pharmaceuticals may lead to overconsumption.

If providers and patients lack full information about available treatments, then direct-to-consumer advertising could allow patients to become more informed about potential treatments for their medical condition. It seems difficult to argue increased use resulting from better-informed customers is entirely inappropriate. As evidence of the potential for this information channel, several studies examining the effect of direct-to-consumer advertising have found spillovers well beyond the advertised drug (for example, Shapiro 2018; Sinkinson and Starc 2019). These studies find net increases in the use of generic drugs—which would be consistent with the advertising efforts increasing the likelihood of a physician visit rather than simply increasing the use of the advertised drug among patients already visiting the physician to receive a prescription.

On the other hand, in an insured population, patients and providers may not be appropriately weighing the value of products if they do not face the marginal cost of purchasing the drug (Shapiro 2022). This could lead to concerns that advertising promotes the use of cost-ineffective medicines. Thus, it is important to carefully examine a full set of effects of the change in use resulting from direct-to-consumer advertising. As one example, Shapiro (2022) examines direct-to-consumer advertising for antidepressant treatments and finds exposure to such advertising increases the labor supply of patients, which implies health gains. Alpert, Lakdawalla, and Sood (2023) found increased adherence for patients already taking a prescription.

The clinical benefits of the drug being advertised will also affect welfare calculations. Some evidence suggests firms may be more likely to advertise drugs providing a

lower clinical benefit (DiStefano et al. 2023). In contrast, Starc and Sinkinson (2019) find meaningful economic benefits from the advertising of statins (drugs to treat high cholesterol), driven in part by the large positive benefits of that class of medications. Indeed, they find the welfare benefits of advertising statins were so large they surpassed overall spending on direct-to-consumer advertising in the market. While a comprehensive analysis of the welfare effects of direct-to-consumer advertising across all products and settings has not yet been offered, and would be difficult, several more focused studies have found net benefits from direct-to-consumer advertising of specific products.

Marketing to Providers

Drug manufacturers also target physicians for advertising through a process known as "detailing." Despite most attention being placed on consumer advertising, most marketing resources are targeted at physicians (Carey, Lieber, and Miller 2021). The vast majority of these interactions involve relatively small costs, like buying a meal, which serves as a testament to the pervasiveness of this activity. Again, the key question is whether visits from pharmaceutical representatives increase the amount of information available to physicians, or only seek to sway physician prescribing behavior through the provision of tangible benefits. This effect is ultimately a difficult empirical question because of a selection problem. After all, we would expect physician marketing to be concentrated on those physicians likely to change their behavior in a way that increases profits (Grennan et al. 2024).

That said, some progress has been made. Carey, Lieber, and Miller (2021) found payments from a manufacturer to a physician lead to a small increase of approximately 2 percent in prescriptions by paid physicians, and an 8 percent increase in patient spending. Back-of-the-envelope calculations show that overall physician payments generate an economically meaningful return on investment for manufacturers. But from a welfare point of view, the important question is not simply whether advertising to physicians increases prescriptions, but whether the additional consumption increases welfare. In turn, this depends on beliefs about why patients previously were not previously gaining access to effective medications and the efficacy of the medications they end up taking.

In a study of the effect of physician payments on the use of anti-coagulant drugs, Agha and Zeltzer (2022) find increases in use both by the affected physician and by the physician's peers. The authors also find greater use across both patients for whom the products could pose a high risk of adverse events and those with a lower risk of such outcomes. This reasonably generates skepticism about, at a minimum, the precision of the information provided by these detailing activities.

Again, an overall welfare calculation of physician-advertising for drugs would be very difficult. At a minimum, the welfare effects for physician detailing are likely also specific to the class of drugs. Grennan et al. (2024) provide a rich examination of welfare effects in the market for statins and find manufacturer-provided meals cause an increase in the use of these products. These newer prescriptions tended to be expensive branded products rather than more inexpensive generic statins. While

this could be worrisome, the under-prescribing of highly effective statins compared to the social optimum was large enough that it outweighed the costs of the higher prices for branded drugs. The effects of such an analysis for a broader set of drugs remains unclear.

It should also be noted that manufacturers coordinate their direct-to-consumer and physician-focused commercialization efforts. Therefore, evaluating the effects of these practices or proposals to regulate them should also seek to understand the potential dynamic effects on other firm behaviors.

Tradeoffs of High Drug Prices in a Heavily Insured Market

The incentives for pharmaceutical innovation currently rely on firms earning positive margins over the medium-run to recoup the research and development and regulatory investments necessary to bring new drugs to market. Understanding the aggregate effects of such policies requires understanding how much welfare is actually decreased by these higher prices. Such a calculation requires carefully considering the impacts of public and private insurance. Well-functioning insurance insulates patients from the direct cost of purchasing medical treatments and therefore could blunt the welfare losses from higher prices (although distributional concerns may remain). Lakdawalla and Sood (2013) estimate quantity changes for a molecule after the loss of a patent swiftly decreases prices. If the previously elevated prices were limiting access to the drug, there should be a corresponding increase in quantity. However, the authors estimate at best a small increase—demonstrating the relatively limited degree to which high prices resulting from greater market power among manufacturers limit consumption in the heavily insured US pharmaceutical market.

That said, these results rely on well-functioning insurance markets, which may not be true in the current market. In particular, as prices rise they could lead to a greater amount of cost-sharing payment requirements from consumers, as will be discussed later in this section. In addition, the results in Lakdawalla and Sood (2013) clearly do not extrapolate to the uninsured. This fact is particularly important to the extent that higher prices can themselves push people out of the market for any health insurance at all—even those who are not interested in purchasing the expensive product. The combination of these factors could decrease overall welfare (Besanko, Dranove, and Garthwaite 2020).

Given the importance of insurance for understanding the welfare implications of high drug prices, it is also important to understand the specific structures of insurance markets. Nearly 40 percent of the United States receives public health insurance coverage. Each of these programs have specific rules that can meaningfully impact patient welfare, Thus, the next three sections focus on three public programs that provide insurance for drugs: Medicaid, the 340(B) Drug Pricing Program, and Medicare Part D. The final subsection focuses on cost-sharing rules for drugs in both public and public health insurance, which can limit the extent to which insurance protects patients from higher drug prices.

Medicaid

Medicaid provides health insurance coverage, including drug insurance, for low-income and disabled Americans. The US government uses two primary methods to ensure Medicaid pays the lowest price in the drug market: (1) drug manufacturers pay state Medicaid systems a rebate off the publicly available list price that is equal to the greater of either 23.1 percent or the largest rebate given to any commercial buyer; and (2) manufacturers must provide additional "inflationary" rebates equal to the difference between the current list price and the list price when the product was first sold. Thus, drug manufacturers with products exhibiting higher list price growth must pay larger Medicaid rebates.

This combination of rebate policies results in Medicaid only paying 35 percent of the average price paid by insurers in the Medicare Part D and the commercial market (CBO 2021). Indeed, many products with high commercial prices are sold to the Medicaid system at effectively no cost. This primarily results from the effect of a larger inflation rebate for products with many years of high list price growth.

Firms will react to Medicaid rebate policies in setting their list prices, depending on factors such as the share of a drug's sales going to Medicaid and how higher list prices will affect revenues in other markets. More specifically, the "best price" rebate increases the cost to manufacturers for providing discounts in the commercial market: in this way, the Medicaid best price rule increased commercial prices (Duggan and Scott Morton 2006). This is effectively an implicit tax on the commercial market to support a lower "on budget" impact of Medicaid. The impact of inflation rebates on commercial list prices is complex. The initial effect is that firms have an incentive to increase the list price at launch. But over time, this provision provides an incentive to limit list price growth (Feng, Hwang, and Maini 2023).

The 340B Drug Pricing Program

The Emergency Medical Treatment and Active Labor Act of 1986 requires hospitals to treat and stabilize all patients regardless of ability to pay. It is part of a system of informal insurance that relies on social norms and explicit regulations governing hospital-based uncompensated care (Garthwaite, Gross, and Notowidigdo 2018). In recognition of the burden this system of informal insurance places on some hospitals, Congress also created programs to subsidize the finances of affected facilities.

One such program, the 340(B) Drug Pricing Program (hereafter 340B), was created in 1992 and allows certain hospitals to purchase pharmaceuticals at approximately the Medicaid best price. Hospitals can then sell these pharmaceuticals in an outpatient setting. When they sell to a Medicaid patient, they must accept the best price as described above. For commercial patients, however, they are free to charge whatever prices they can negotiate. Similarly, when hospitals sell to patients who are uninsured (or who face large amounts of cost sharing) hospitals can, but are not required to, set lower prices that increase access. A large fraction of hospitals do not pass those discounts onto their uninsured or underinsured patients (GAO 2018).

Initially, 340B was a small program that stringently limited the number of qualifying hospitals and the number of pharmacies per hospital. These limitations have

since eased. From 2010 to 2023, the number of contract pharmacies participating in the program increased from 1,300 to 33,000—roughly half of all free-standing pharmacies in the country. From 2010 to 2021, the dollar value of products in the program (valued at the list price) increased from $6.6 billion to $43.9 billion (Sachs and Varcie 2024). This spending is not evenly divided across products: oncology products, anti-infectives, and immunosuppressants account for 70 percent of all spending.

Compared to other drug insurance programs, 340B is relatively understudied. However, as the program grows in size, it can have market-wide implications. At a minimum, to the extent 340B reduces the available revenues for innovative drugs, it will decrease the financial returns available for successful products and diminish incentives to invest in new product development.

Moreover, 340B has a similar effect on commercial drug prices as the Medicaid program described above. The program also can have additional effects. First, in some settings manufacturers still pay rebates to plan sponsors on products covered by 340B. This further reduces revenues per product and provides an incentive to reduce the optimal rebate provided to commercial plans. Alternatively, for some purchases drug manufacturers do not pay rebates for products purchased through 340(B). This results in plan sponsors paying list prices for a higher fraction of their prescription purchases. Gray (2023) finds that growth in the size of 340B increases commercial health insurance premiums. Most existing research on 340(B) focuses on the types of hospitals receiving the product and how they pass along (or do not) the discounts they receive. More research is needed to understand the marketwide implications of this growing program.

Medicare Part D

In 2006, Medicare Part D was created as a retail drug insurance program for the elderly. Part D is heavily subsidized by the government, yet operated by private firms who were responsible for developing formularies, creating networks of pharmacies, and administering all aspects of the insurance program. Prices for these drugs are negotiated in the same manner as the overall commercial market and have tended to be about the same as commercial market prices.

At its creation, Part D included both a subsidy and a reinsurance component. After enrollees spent a certain amount of their own funds out-of-pocket they enter the "catastrophic" region of the program where the government assumed responsibility for 80 percent of the additional spending. The plan was still required to pay for 15 percent of additional spending, and patients without any form of supplemental insurance were required to pay 5 percent of list prices with no cap on overall spending.

Under such a system, both Part D plans and manufacturers—that is both sides of the drug price negotiation in that program—have an incentive for list prices to rise quickly, because it would lead to higher out-of-pocket spending and more quickly move high-cost patients into the catastrophic region of the program (that is, the region where the government assumes most of the prescription drug costs).

Ippolito and Levy (2023) found that products with a larger exposure to Medicare Part D had greater differences between their list and net prices—and that these differences increased after Part D made manufacturers responsible for more of the cost for expensive patients. This estimated effect based on exposure to Medicare grows in time markedly following 2012—which is also the time of the market-wide divergence in list and net revenue shown in Figure 2.

As part of the 2022 Inflation Reduction Act, the structure of the catastrophic region of the program was changed so the majority of spending was now paid by the plan sponsor. In addition, the portion previously paid by the patient was eliminated, which greatly reduced patient-level cost sharing. The intention of this policy change was to shift more burden onto plan sponsors in order to provide stronger incentives for negotiation and alleviate the cost-sharing burdens on enrollees. Initial evidence suggests that the change may also increase the subsidy to plans from the government and require higher premiums.

A further provision of the Inflation Reduction Act granted authority to the Centers for Medicare and Medicaid Services (CMS) to demand lower prices from drug manufacturers under certain conditions. CMS can select drugs for this program 9 years after first release for traditional small molecule products and 13 years for biologics. This time period is not extended if firms receive approvals for new indications beyond their initial approval.

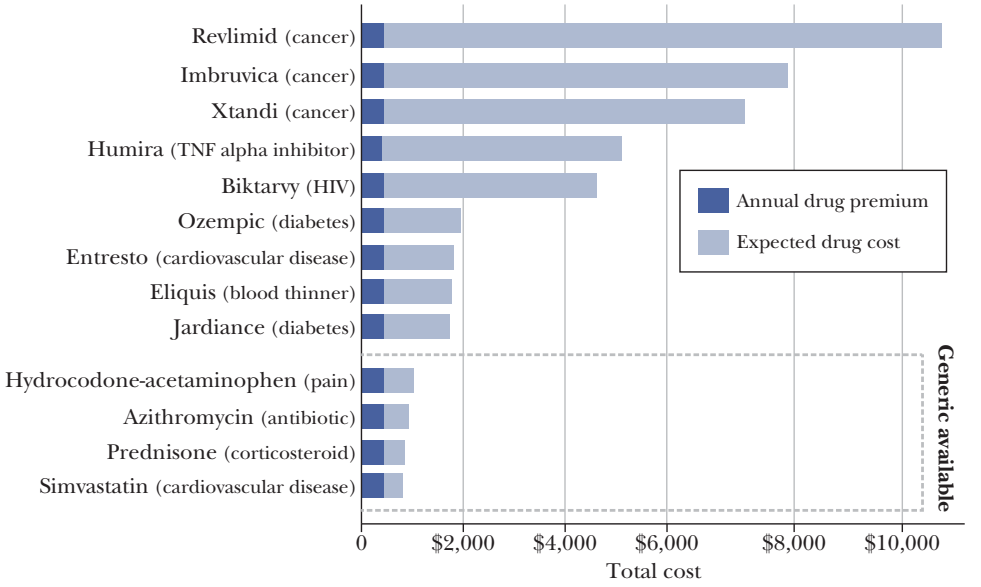
The Medicare Part D provisions of the 2022 legislation are still being implemented, and they are a subject of ongoing controversy and research. In particular, there is debate about its possible magnitude on innovation, which is due in part to uncertainty about the size of the discounts that will be demanded and how drugs will be selected for the program. In addition, the reliance on first date of entry to the market and the lack of an extension for new uses will influence both the optimal launch strategy of firms and the willingness to invest in research demonstrating additional indications for existing products. Finally, if the law does impose higher costs on Part D plan sponsors, then those firms would have a reduced incentive to participate in the program.

Strategic Cost Sharing in Drug Insurance Markets

Many medications taken by patients are treatments of chronic conditions with largely certain annual consumption and resulting cost sharing. To the degree that a patient's health status determines their pharmaceutical usage, the existence of high cost sharing creates a higher effective health insurance premium for the year based on an enrollee's health status. This essentially reintroduces medical underwriting to the insurance market for patients with conditions requiring expensive pharmaceutical treatments. Notably, patients requiring intensive use of hospital services, for example, are not subject to similar implicit medical underwriting.

This system transfers money from sick enrollees to plan sponsors that can either keep it as profits or decrease premiums in order to gain market share. This strategy appears to be employed, albeit in different ways, by many Medicare Part D insurers and commercial firms. As noted above, the structure of Medicare Part D causes

Figure 3
The Median Medicare Plan D Plan: Some Examples of Total Cost by Drug



Source: Garthwaite and Starc (2023).

increases in out-of-pocket spending to shift the payment responsibilities from the insurer and the manufacturer to the government. To illustrate this point, Figure 3 contains the median annual cost (that is, insurance premium and cost-sharing payments) for a Medicare Part D enrollee taking specific chronic prescription drugs. For certain drugs, the annual costs exceed $8,000.

In the commercial market, manufacturers have been able to diminish the effect of these cost-sharing payments through various assistance programs. Perhaps the most widely discussed of these are copayment coupons—which involve the manufacturer paying cost sharing on behalf of the consumer. Such payments are allowed in the commercial market, but are a violation of federal law for government programs such as Medicare and Medicaid.

For the drug company, the strategic rationale of such cost-sharing assistance is clear. Consider a high-priced drug with a 20 percent coinsurance rate paid by the patient. In that case, a copayment coupon effectively amounts to a manufacturer offering an additional 20 percent discount. These cost-sharing payments still likely decrease premiums for healthy patients, but the initial transfer is coming from manufacturers rather than sick patients. The ultimate incidence in these markets is unclear.

The broader implications of cost-sharing assistance vary by the type of competition a product faces. Dafny, Ody, and Schmitt (2017) examine the situation where a manufacturer introduces a copayment coupon for a product facing generic

competition. However, there is no clear evidence these coupons provide increased access to a particular molecule. Instead, they simply shift market share to the branded version of the product and increase profits for the original manufacturer.

Manufacturers also provide copayment assistance for products that face high cost sharing but have no generic substitutes. Such coupons are estimated to increase the net price of such products by approximately 8 percent (Dafny, Ho, and Kong 2024). This occurs because copayment assistance blunts the ability of the commercial insurer to use cost-sharing payments during negotiations. Unlike in the case of generic competition, these coupons provide increased access to specific molecules. Therefore, the welfare effects hinge on the degree of substitutability among the branded products in the market and heterogeneity in the treatment effects across products.

Cost-sharing assistance from drug manufacturers does not exist in a vacuum. Pharmacy benefit managers and insurance plan sponsors may attempt to blunt the effect of these strategies, with perhaps the most extreme reaction being an increase in the number of drugs simply not covered. Because uncovered drugs are effectively offered at a coinsurance rate of 100 percent, it is not profitable for manufacturers to offer a coupon. As of 2023, each of the three major pharmacy benefit manager firms (Optum, Caremark, and Express Scripts) excluded approximately 600 products from their formulary (Fein 2023). Pharmacy benefit managers have also created systems that limit the ability of coupons to cover the broader out-of-pocket obligations such as deductibles and annual out of pocket limits.

The literature on the effect of cost-sharing assistance by drug manufacturers on other aspects of negotiation between insurers and patients also provides some insight into the likely effects of policies to limit cost sharing market-wide such as passed caps on total out-of-pocket spending for insulin in the Medicare Part D program, included as a provision in the Inflation Reduction Act of 2022. Depriving insurers of the access to provide higher cost sharing on at least some products in a therapeutic class reduces the ability to use cost sharing as a negotiating tool.

Negotiations at the Time of Transition from Market Exclusivity

Intellectual property protections for innovative firms are meant to be time-limited. After this period has expired, optimal policy requires robust competition between the incumbent firm and new entrants. Given this threat of competition, brand-name drug manufacturers have a strong incentive to seek out ways to extend their period of exclusivity in part through a more expansive use of tools such as patents and trade secrets.

Given the requirement that patents are narrow and specific to a particular invention, modern complex drugs are often covered by a wide range of patents. At a high level we can think about intellectual property related to either: (1) the main molecule; (2) the production process for that molecule; and (3) novel additional uses for that molecule. The last two categories, in particular, may be updated with

additional patents over time. There are concerns this can create a "patent thicket"—that is, numerous overlapping patents that raise the costs of entry with no clear end date. Indeed, some critics have gone as far as to suggest that each branded drug should be limited to a single patent (Feldman 2019).

While some firms likely engage in entry-deterring "thicketing" strategies, the mere existence of numerous patents is not, on its own, evidence of nefarious intentions. Beyond the shifting complexity of production, pharmaceutical products are increasingly used to treat multiple conditions. Given the potentially great benefits to society from firms developing new uses for existing products, policies that encourage firms that invest in research and development for such uses could increase welfare. Therefore, concerns over intellectual property should concentrate on the validity of underlying patents, rather than simply counting their number.

Even if existing intellectual property protections are appropriately granted, a well-functioning system still requires a system of robust post-exclusivity competition. For "small molecule" products, a generic competitor can manufacture an exactly bioequivalent product. In this setting, the entry of generic producers when patent terms expire results in large price reductions as long as the overall market is large enough for a number of entrants to produce at a minimum efficient scale. If the market is not large compared to minimum efficient scale, then it may have room for only a few or even only one manufacturer. In recent years, several firms appear to have recognized the pricing power available to manufacturers of generic products with sufficiently small potential markets. Concerns about the market for small market generics are growing as we see more entry of precision products treating very small patient populations (Chandra, Garthwaite, and Stern 2017). Finding ways to reduce entry costs and decrease delays for generics could leading to greater increases in consumer benefits (Starc and Wollman forthcoming).

Many recent innovations in drug markets are "large molecule" biologic products. In this setting, new entrants currently lack the ability to create exact replicas, and so when exclusivity expires, the incumbent brand-name drug firms face competition from a biosimilar. These products are supported by clinical trial evidence that they perform in a clinically similar manner to the reference product, but are not truly bioequivalent. Markets for biosimilars have largely struggled to obtain price decreases anywhere close to those for generic drugs. Given that biosimilars are not bioequivalent, current policies do not allow for automatic substitution of the reference product with the biosimilar at the pharmacy counter. Indeed, patients who are currently using a brand-name biologic product may be unwilling to switch to a competing biosimilar at almost any price. As a result, entrants with large molecule drugs must commercialize the product through the types of activities described above.

The structure of rebate contracts can also potentially create barriers to entry as well. For example, a rebate contract for an incumbent product may provide rebates only if the pharmacy benefit manager does not include a rival entrant in the list of drugs available in the formulary, which is sometimes referred to as a rebate "wall" or "trap." A pharmacy benefit manager may have an incentive to solicit a rebate

from a manufacturer for exclusive formulary placement. Although individual pharmacy benefit managers could benefit from such a contract, a pattern of these rival-referencing contracts can lead to a situation in which potential manufacturers of biosimilars may choose not to attempt to create products in the first place.

As these examples illustrate, the incentives for innovation in pharmaceuticals are strongly dictated by barriers to new products and are therefore heavily influenced by the specific details of intellectual property. When a drug receives additional patents beyond the active ingredient, have such patents been appropriately examined by the US Patent and Trademark Office? What rules govern small-market drugs where entry at minimum sustainable size may be impractical? What rules govern the entry of biosimilars that are not bioequivalents but can provide important value creating competition? What rules govern permissible contract structures for formularies, rebates, and patient cost sharing? It should not be surprising that for any given drug, well-established incumbent firms will always be on the lookout for ways to extend their market exclusivity, and potential but currently nonexistent entrants may be poorly represented in drawing up these rules. Hoping for self-regulation to lead to an optimal form of regulation is foolhardy behavior. These questions reflect the overall tradeoffs present in a system using market-based incentives to promote and maintain welfare-creating innovation.

References

Acemolgu, Daron, and Joshua Linn. 2004. "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry." *Quarterly Journal of Economics* 119 (3): 1049–90.

Agha, Leila, Soomi Kim, Danielle Li. 2022. "Insurance Design and Pharmaceutical Innovation." *American Economic Review: Insights* 4 (2): 191–208.

Agha Leila, and Dan Zeltzer. 2022. "Drug Diffusion through Peer Networks: The Influence of Industry Payments." *American Economic Journal: Economic Policy* 14 (2): 1–33.

Alpert, Abby, Darius Lakdawalla, and Neeraj Sood. 2023. "Prescription Durg Advertising and Drug Utilization: The Role of Medicare Part D." *J Public Econ*. doi: 10.1016/j.jpubeco.2023.

Angelis, Aris, Roman Polyakov, Olivier J. Wouters, Els Torreele, and Martin McKee. 2023. "High Drug Prices Are Not Justified by Industry's Spending on Research and Development." *British Medical Journal* 380: e071710.

Bagwell, Kyle. 2007. "The Economic Analysis of Advertising." In *Handbook of Industrial Organization*, Vol. 3, edited by M. Armstrong and R. Porter, 1701–1844. North-Holland.

Berger, Benjamin, Amitabh Chandra, and Craig Garthwaite. 2021. "Regulatory Approval and Expanded Market Size." NBER Working Paper 28889.

Besanko, David, David Dranove, and Craig Garthwaite. 2020. "Insurance Access and Demand Response: Pricing and Welfare Implications." *Journal of Health Economics* 73: 102329.

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

Budish, Eric, Benjamin N. Roin, and Heidi Williams. 2015. "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials." *American Economic Review* 105 (7): 2044–85.

Buxbaum, Jason D., Michael E. Chernew, A. Mark Fendrick, and David M. Cutler. 2020. "Contributions of Public Health, Pharmaceuticals, and Other Medical Care to US Life Expectancy Changes, 1990–2015." *Health Affairs* 39 (9): 1546–56.

Carey, Colleen, Ethan M. J. Lieber, and Sarah Miller. 2021. "Drug Firms' Payments and Physicians' Prescribing Behavior in Medicare Part D." *Journal of Public Economics* 197: 104402.

Chandra, Amitabh, John Drum, Michael Daly, Henry Mirsberger, Samuel Spare, Ulrich Neumann, Silas Martin, and Noam Kirson. 2024. "Comprehensive Measurement of Biopharmaceutical R&D Investment." *Nature Reviews Drug Discovery* 23: 652–53.

Chandra, Amitabh, Craig Garthwaite, and Ariel Dora Stern. 2017. "Characterizing the Drug Development Pipeline for Precision Medicine." NBER Working Paper 24026.

Congressional Budget Office (CBO). 2021. *A Comparison of Brand-Name Drug Prices among Selected Federal Programs*. Congressional Budget Office.

Dafny, Leemore, Kate Ho, Edward Kong. 2024. "How Do Copayment Coupons Affect Branded Drug Prices and Quantities Purchased?" *American Economic Journal: Economic Policy* 16 (3): 314–46.

Dafny, Leemore, Christopher Ody, and Matt Schmitt. 2017. "When Discounts Raise Costs: The Effect of Copay Coupons on Generic Utilization." *American Economic Journal: Economic Policy* 9 (2): 91–123.

Dean, Emma B., Josh Feng, and Luca Maini. 2024. "Stocking Under the Influence: Spillovers from Commercial Drug Coverage to Medicare Utilization." https://static1.squarespace.com/static/5b3660f9b98a78542ce0faa9/t/66b0e92b2cf4e015d36c2a2c/1722870060624/StockingUnderTheInfluence_07-29-24.pdf.

DiStefano Michael J., Jenny M. Markell, Caroline C. Doherty, G. Caleb Alexander, and Gerard F. Anderson. 2023. "Association between Drug Characteristics and Manufacturer Spending on Direct-to-Consumer Advertising." *JAMA* 329 (5): 386–92.

Dranove, David, Craig Garthwaite, Christopher Heard, and Bingxiao Wu. 2022. "The Economics of Medical Procedure Innovation." *Journal of Health Economics* 81: 102549.

Dranove, David, Craig Garthwaite, Manuel Hermosilla. 2022. "Does Consumer Demand Pull Scientifically Novel Drug Innovation?" *RAND Journal of Economics* 53 (3): 590–638.

Dubois, Pierre, Olivier de Mouzon, Fiona Scott-Morton, and Paul Seabright. 2015. "Market Size and Pharmaceutical Innovation" *RAND Journal of Economics* 46 (4): 844–71.

Duggan, Mark, and Fiona M. Scott. 2006. "The Distortionary Effects of Government Procurement: Evidence from Medicaid Prescription Drug Purchasing." *Quarterly Journal of Economics* 121 (1): 1–30.

Eliason, Paul J., Benjamin Heebsh, Ryan C. McDevitt, and James W. Roberts. 2020. "How Acquisitions Affect Firm Behavior and Performance: Evidence from the Dialysis Industry." *Quarterly Journal of Economics* 135 (1): 221–67.

Fein, Adam J. 2023. "The Big Three PBMs' 2023 Formulary Exclusions: Observations on Insulin, Humira, and Biosimilars." Drug Channels, January 10. https://www.drugchannels.net/2023/01/the-big-three-pbms-2023-formulary.html.

Feldman, Robin. 2019. "'One-and-Done' for New Drugs Could Cut Patent Thickets and Boost Generic Competition." *STAT News*, Feb 11. https://www.statnews.com/2019/02/11/drug-patent-protection-one-done/.

Feng, Josh, Thomas Hwang, and Luca Maini. 2023. "Profiting from Most-Favored-Customer Procurement Rules: Evidence from Medicaid." *American Economic Journal: Economic Policy* 15 (2): 166–97.

Finkelstein, Amy. 2004. "Static and Dynamic Effects of Health Policy: Evidence from the Vaccine Industry." *Quarterly Journal of Economics* 119 (2): 527–64.

Frakes Michael D., and Melissa F. Wasserman. 2023. "Investing in Ex Ante Regulation: Evidence from Pharmaceutical Patent Examination." *American Economic Journal: Economic Policy* 15 (3): 151–83.

Gallup. 2024. "Business and Industry Sector Ratings." https://news.gallup.com/poll/12748/business-industry-sector-ratings.aspx (accessed April 6, 2025).

Garthwaite, Craig L. 2012. "The Economic Benefits of Pharmaceutical Innovations: The Case of Cox-2 Inhibitors." *American Economic Journal: Applied Economics* 4 (3): 116–37.

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

Garthwaite, Craig, Rebecca Sachs, and Ariel Dora Stern. 2022. "Which Marks (Don't) Drive Pharmaceutical Innovation? Evidence from U.S. Medicaid Expansions." NBER Working Paper 28755.

Garthwaite, Craig, and Amanda Starc. 2023. "Why Drug Pricing Reform Is Complicated: A Primer and Policy Guide to Pharmaceutical Prices in the US." In *Building a More Resilient US Economy*, edited by

Melissa S. Kearney, Justin Schardin, and Luke Pardue, 74–128. Aspen Institute.

Government Accountability Office (GAO). 2018. *Federal Oversight of Compliance at 340B Contract Pharmacies Needs Improvement*. Government Accountability Office.

Gray, Charles. 2023. "The Incidence of the 340B Program: THe Effects of Contract Pharmacies on Part D Premiums and Reimbursements." Working Paper.

Grennan, Matthew, Kyle R. Myers, Ashley Swanson, and Aaron Chatterji. 2024. "No Free Lunch? Welfare Analysis of Firms Selling through Expert Intermediaries." https://doi.org/10.1093/restud/rdae090.

Grennan, Matthew, and Robert J. Town. 2020. "Regulating Innovation with Uncertain Quality: Information, Risk, and Access in Medical Devices." *American Economic Review* 110 (1): 120–61.

Ippolito, Benedic N., and Joseph F. Levy. 2023. "The Influence of Medicare Part D on the List Pricing of Brand Drugs." *Health Services Research* 58 (4): 948–52.

IQVIA Institute. 2024. *The Use of Medicines in the U.S. 2024: Usage and Spending Trends and Outlook to 2028*. IQVIA Institute.

Jena, Anupam B., John E. Calfee, Edward C. Mansley, and Tomas J. Philipson. 2009. "'Me-Too' Innovation in Pharmaceutical Markets." *Forum for Health and Economic Policy* 12 (1): 5.

Jena, Anupam B., and Tomas Philipson. 2007. "Cost-Effectiveness as a Price Control." *Health Affairs* 26 (3): 696–703.

Kakani, Pragya, Michael Chernew, and Amitabh Chandra. 2022. "The Contribution of Price Growth to Pharmaceutical Revenue Growth in the United States: Evidence from Medicines Sold in Retail Pharmacies." *Journal of Health Politics, Policy, and Law* 47 (6): 629–48.

Krieger, Joshua, Danielle Li, and Dimitris Papanikolaou. 2022. "Missing Novelty in Drug Development." *Review of Financial Studies* 35 (2): 636–79.

Lakdawalla, Darius N. 2018. "Economics of the Pharmaceutical Industry." *Journal of Economic Literature* 56 (2): 397–449.

Lakdawalla, Darius, Anup Malani, and Julian Reif. 2017. "The Insurance Value of Medical Innovation." *Journal of Public Economics* 145: 94–102.

Lakdawalla, Darius, and Neeraj Sood. 2013. "Health Insurance as a Two-Part Pricing Contract." *Journal of Public Economics* 102: 1–12.

Mulligan, Casey B. 2023. "Ending Pay for PBM Performance: Consequences for Prescription Drug Prices, Utilization, and Government Spending." NBER Working Paper 31667.

Peltzman, Sam. 1973. "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments." *Journal of Political Economy* 81 (5): 1049–91.

Sachs, Rebecca, and Joshua Varcie. 2024. "Spending in the 340B Drug Pricing Program, 2010 to 2021." Presentation, 13th Annual Conference of the American Society of Health Economists, Congressional Budget Office, June 17, 2024. https://www.cbo.gov/publication/60339.

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. 2022. "Promoting Wellness or Waste? Evidence from Antidepressant Advertising." *American Economic Journal: Microeconomics* 14 (2): 439–77.

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

Sparks, Grace, Ashley Kirzinger, Alex Montero, Isabelle Valdes, and Liz Hamel. 2024. "Public Opinion on Prescription Drugs and Their Prices." Kaiser Family Foundation, October 4. https://www.kff.org/health-costs/poll-finding/public-opinion-on-prescription-drugs-and-their-prices/.

Starc, Amanda. 2023. "Manufacturer Spending on Direct-to-Consumer Advertising for Pharmaceutical Products." *JAMA* 329 (5): 371–73.

Starc, Amanda, and Thomas G. Wollmann. Forthcoming. "Does Entry Remedy Collusion? Evidence from the Generic Prescription Drug Cartel." *American Economic Review*.

Yeung, Kai, Lisa Bloudeck, Yao Ding, and Sean D. Sullivan. 2022. "Value-Based Pricing of US Prescription Drugs: Estimated Savings Using Reports from the Institute for Clinical and Economic Review." *JAMA Health Forum* 3 (12): e224631.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 27–52

Patents, Innovation, and Competition in Pharmaceuticals: The Hatch-Waxman Act After 40 Years

C. Scott Hemphill and Bhaven N. Sampat

Pharmaceuticals are the classic example of how patents work to balance dynamic and static efficiency. Patents provide exclusivity for a firm to sell an innovative "branded" drug at a price well above marginal cost, thus promoting dynamic efficiency by encouraging investments in innovation, resulting in new drugs that improve health. But the higher prices during this time represent a source of lost welfare to purchasers. Moreover, uninsured or underinsured patients may be priced out of the treatment and suffer worse health outcomes. At some point the patent protection expires, and competing "generic" drugs enter the market at a low price, a source of static efficiency. The central issue in pharmaceuticals, as a matter of economic inquiry and regulatory design, is how to balance the dynamic benefits of new drugs against the static benefits of low prices for existing drugs.

In the United States, that balance is set by the Drug Price Competition and Patent Term Restoration Act of 1984, commonly known as the Hatch-Waxman Act. The law represents a compromise, as reflected in its compound name.

First, a set of "price competition" provisions seek to facilitate quick generic entry once patents no longer stand in the way. They pave the way for generics to enter at a lower price—the more generics, the lower the price (Olson and Wendling 2018)—and

■ *C. Scott Hemphill is the Moses H. Grossman Professor of Law, New York University School of Law, New York City, New York. Bhaven N. Sampat is Professor, School for the Future of Innovation in Society and School of Public Affairs, Consortium for Science, Policy and Outcomes Professor, Arizona State University, Washington, DC. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Sampat is the corresponding author at bhaven.sampat@asu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241423. The authors have received financial support from interested parties.

typically capture the great majority of the market very quickly (Buttorff, Xu, and Joyce 2020). Rapid market penetration is promoted by laws in nearly every state, which require or permit a pharmacist to fill a prescription with a generic substitute where available. Private and government payers further encourage substitution by lowering a patient's out-of-pocket cost when a generic product is used. Generic medicines saved $3.1 trillion for drug purchasers over the decade from 2014 through 2023, and $445 billion in 2023 alone (AAM 2024). These provisions also provide a means to handle frequent, contentious, and high-stakes disagreements between the branded and generic firms about whether particular patents are actually valid and infringed, and thereby block generic entry.

Meanwhile, the "patent term restoration" provisions extend the patent term to allow branded drug makers to recoup part of the time spent during clinical trials and review by the Food and Drug Administration (FDA). Restoration aims to augment the incentive to innovate. This matters because newly patented drugs play an important role in helping people to live longer, healthier lives (Cutler, Deaton, and Lleras-Muney 2006; Buxbaum et al. 2020; Lichtenberg 2022). From early-stage research and development investment to government-mandated clinical trials, bringing a new branded drug to market is very costly. Recent estimates place the cost above $1.5 billion per approved drug (DiMasi, Grabowski, and Hansen 2016; Wouters, McKee, and Luyten 2020)—a figure that accounts for the fact that most drug development efforts are a failure. A half-century of cross-industry economic research furnishes empirical evidence that in the pharmaceutical industry, patents are a uniquely important source of appropriability (Taylor and Silbertson 1973; Mansfield 1986; Levin et al. 1987; Cohen, Nelson, and Walsh 2000; Mezzanoti and Simcoe 2023).

The effectiveness of both sets of provisions is open to question. Most obviously, high prices for branded drugs are a subject of public and political attention. Observers have raised concerns that patent-holding firms have erected barriers to generic competition, thereby forestalling generic entry and creating static inefficiency. As for dynamic efficiency, the other side of the ledger, partial patent term restoration leaves intact a distortion whereby drugs with longer clinical trials receive shorter periods of exclusivity.

Measured against its goal of establishing a pathway for generic entry, the Hatch-Waxman Act has been a success. At the same time, setting the right balance between dynamic and static efficiency remains a pressing question. The 40th anniversary of the law presents an opportunity to assess its successes and failures in promoting innovation and competition. In this essay, we begin by reviewing the Act's origins and key features. We then present evidence on how well the law has encouraged competition and rewarded innovation in pharmaceutical markets. On the competition side, we show how the Act creates incentives for branded firms to accumulate patents and generics to challenge them, and various strategies deployed by branded firms to delay generic entry. On the innovation side, we characterize the prevalence and extent of patent restoration. The net result of the Hatch-Waxman compromise is a convoluted and expensive approach to balancing

innovation and competition—one that is unlikely to align social and private rewards to innovation. Finally, we consider various avenues for reform, ranging from tweaks of Hatch-Waxman to alternative policy approaches that hold promise in achieving these goals at lower cost.

Background

Before the Hatch-Waxman Act

In evaluating the evidence that has accumulated in the four decades since the Hatch-Waxman Act was passed, it is useful to reprise how we got here and the problems to which the Act was addressed. To do so, we must harken back to the birth of the modern research-intensive pharmaceutical industry at the end of World War II.

A government-led medical research effort during that war, enlisting universities and firms across the country, helped support the mass production of penicillin, malaria treatments, a range of vaccines, steroids, blood substitutes, and many other treatments that were crucial for the Allied victory (Gross and Sampat 2023). During the war, most medical research contracts prohibited researchers from obtaining patents covering the fruits of government-funded research (Sampat 2020; Gross and Sampat 2023). Instead, a guaranteed market and wartime subsidies made the effort worthwhile for participating firms (Temin 1980). Thanks to this arrangement, firms developed research capabilities, identified drug innovation as a source of high profits, and built research and development programs (Temin 1979; Temin 1980; Cockburn et al. 1999).

After World War II, the US government left drug development to firms, confining its funding to basic medical research at universities through the National Institutes of Health (Sampat 2023). In a departure from wartime cooperation and knowledge-sharing, "innovator" drug makers increasingly relied on patents as a vehicle to limit competition and provide profits (Temin 1980). In turn, patent protection increased the profitability of other investments such as advertising (Temin 1980). Fueled by new technological opportunities created by the war, and increased appropriability secured by patents, the 1950s marked what is sometimes called the golden era of pharmaceutical innovation.

With the rise of heavy marketing of patent-protected drugs, some members of the medical community raised concerns about over-prescription of drugs with limited efficacy (Podolsky 2015). In response, Congress passed a 1962 law that became known as the Kefauver Act. The law required evidence of efficacy as a condition for regulatory approval by the Food and Drug Administration, a new responsibility for an agency that had been focused primarily on safety up to that point. Specifically, a requirement for "adequate and well-controlled investigations" was institutionalized in rules requiring clinical trials. Senator Estes Kefauver (D-Tennessee), the driving force behind the law, would have gone further by prohibiting patents on molecular modifications that did not improve efficacy (Greene and Podolsky 2012). Kefauver

also proposed compulsory licensing of all patented drugs after an exclusivity period of just three years, with royalties capped at 8 percent of sales.[1]

The generic drug industry emerged as the byproduct of an FDA decision to apply the new efficacy rules to pre–Kefauver Act drugs. A retroactive review called the Drug Efficacy Study Implementation applied Kefauver Act efficacy standards to all drugs approved between 1938 and 1962. For drugs deemed efficacious, manufacturers could continue to market under their pre-1962 regulatory approval. Drugs that failed the test could no longer be marketed. The question arose of how to treat firms wishing to sell copies of drugs deemed effective once they came off-patent. Requiring a new drug application supported by clinical trials for drugs that were off-patent and already deemed efficacious "seemed both economically wasteful and potentially unethical" (Greene 2014, p. 66). Accordingly, in 1969, the FDA created a shortcut—approval by means of an "abbreviated new drug application" (ANDA), which removed the need for clinical trials. This less expensive approach provided an entry pathway for producers of off-patent drugs, who would come to be known as generic producers.

Many generic firms "cut their teeth" developing pre-1962 drugs whose patents had expired (Greene 2014). As this sector grew to be a source of low-cost drugs, other policies supported further growth. In the late 1970s, most states passed laws that (as mentioned earlier) encouraged substitution with generics (Song and Barthold 2018).

By the late 1970s, as patents on post-1962 drugs began to expire, the nascent generic pharmaceutical industry started lobbying to extend the abbreviated new drug application process to these drugs as well (Greene 2014). Absent such a pathway, a generic drug maker was forced to file a new drug application as if the drug had just been invented. The predictable result was that many (reportedly, more than 100) post-1962 drugs had no generic competition even after their patents expired.

The Kefauver Act also had major effects on innovator drug firms. The introduction of costly clinical trials increased the dependence of such firms on patents: without the promise of some exclusivity, firms were reluctant to undertake that expense. Patenting practice, then as now, was to file a patent application at the time of discovery, with patent issuance several years thereafter. As a result, a substantial part of the patent term was consumed by clinical trials and regulatory review. For example, if a patent was issued in 1970 but trials were not complete until 1975, only 12 years of patent term were left out of the statutory 17. (In 1995, the US patent term was changed from 17 years from the date of issuance to 20 years from the date of application filing.[2]) Branded firms argued that shorter effective patent protection was responsible for a decline in new drug introductions. Lost innovation was

[1] Congressional Record. 1961. 87th Congress, 1st Session, Vol. 107, Part 5 (April 12): 5369.

[2] The change was made to comply with international obligations under the Trade Related Intellectual Property Rights (TRIPS) Agreement.

arguably more important than the clinical benefits lost while waiting for government approval, itself the subject of a significant literature (Peltzman 1973).

Main Features of the Hatch-Waxman Act

The Hatch-Waxman Act embodied a compromise. To promote competition, it created a pathway for accelerating approval of generic versions of post-1962 drugs. The 1984 Act set up the long-sought pathway, still used today, under which a generic drug maker can file an abbreviated new drug application relying on clinical data about the existing branded drug. To obtain FDA approval, a drug maker must demonstrate "bioequivalence," meaning that the generic product uses the same active ingredient and is absorbed by the body at the same rate and to the same extent as the brand-name drug. Bioequivalence is also typically necessary for pharmacist substitution under the state laws discussed above.

To address the problem of patent term lost to clinical trials and regulatory review, Hatch-Waxman provided for patent extensions. The extension is available for drugs with a novel active ingredient. Firms marketing a drug can choose one patent per drug for extension, and the extension is added back on to the expiration date of the chosen patent. Restoration is partial: all of the regulatory review period, but only one-half of the testing phase.[3] Any time lost to lack of due diligence by the applicant is subtracted. Restoration of patent time is subject to various caps, put in place at the behest of generic firms and others who argued that patent terms were already too long. A patent can be extended a maximum of five years,[4] and the resulting term of that patent cannot exceed 14 years from drug approval.

In addition to these provisions for generic drug entry and patent term restoration, Hatch-Waxman provided for new regulatory exclusivity. For example, "new chemical entities" are protected from generic competition for five years after approval. A new chemical entity is a drug product containing no active ingredient that has previously been approved. This five-year benefit—which Engelberg (1999) suggests was key to the Hatch-Waxman compromise—was meant to guarantee each such drug a minimum amount of exclusivity without regard to patent protection. This lower-bound protection is important where patent protection is short-lived (near expiration by the time of drug approval, despite restoration) or unavailable—for example, because the underlying innovation is already well-known (Roin 2009).

Generic Entry Prior to Patent Expiration

One obvious strategy for a generic firm considering entry is to seek approval of its abbreviated new drug application once patent expiration occurs. However, matters

[3] In cases where the patent issues after the regulatory review period begins, only the portion of that period after issue is counted, and similarly for the testing phase. Patents issued after a drug's approval cannot be extended.

[4] "Pipeline" drugs—drugs that were in trials but not yet approved at the passage of Hatch-Waxman—had a cap of two years.

are not so simple. The patent might be invalid or easily invented around. Moreover, the branded firm may have multiple patents on the drug, which are applied for and expire at different times, resulting in a lengthy period of exclusivity if the generic firm waits until the last patent expires.

Branded firms have a large incentive to use these patents (and other methods) to delay generic entry. After all, later entry extends the period of high prices, increasing branded profits while reducing the static welfare of purchasers. To illustrate, suppose a branded drug has $1 billion in annual sales, and generics achieve 80 percent generic penetration in the first year of entry, during which they sell at a 40 percent discount to the brand.[5] Under these and other assumptions, a one-year delay of generic entry represents a transfer of more than $300 million from purchasers to the branded firm (that is, $1 billion $\times$ 80% $\times$ 40% $=$ $320 million). Even one month of delay can be worth tens of millions of dollars.[6]

With these concerns in mind, the 1984 law provides a three-step process governing patenting and generic entry, including a means for generics to enter prior to patent expiration.

Step One: Patent listing in the Orange Book. As part of its new drug application, a branded drug maker is obliged to report to the Food and Drug Administration certain patents that apply to its approved drug. Most obviously, the drug maker reports patents that claim the drug's active ingredient. In addition, "drug product" patents claiming the formulation or composition and "method" patents claiming a novel use of the product are also listed. The drug maker has a responsibility to submit and describe these patents. The FDA does not evaluate these claims, viewing its role as purely ministerial. The FDA lists these (unaudited) patents in the so-called Orange Book, so named because its first edition was published with an orange cover. The official name is "Approved Drug Products with Therapeutic Equivalence Evaluations," and as the longer title suggests, the document describes the set of generic drugs that are readily substitutable for a branded product. The Orange Book was first published in October 1980; patent listings were added in 1985.

Step Two: Paragraph IV certification. Often, a generic drug maker believes that a patent listed in the Orange Book should not block generic entry because the

[5] These assumptions are conservative. Conrad and Lutter (2019) report a generic-to-brand price ratio of 0.61 with one generic entrant; the ratio is lower with more entrants. Buttorff, Xu, and Joyce (2020) report a typical generic penetration rate of at least 95 percent. The calculation in text further assumes that quantity is unchanged after low-price generic entry occurs. In fact, price elasticity is typically very low. The small consumer response to a price change has multiple sources, including the separation between a physician's decision to prescribe a drug and the decision by the government and private insurers to pay for it. Gatwood et al. (2014), consistent with other studies, reports very small elasticity—between 0 and 0.157.

[6] For hypertension drugs sold between 2000 and 2008, Branstetter, Chatterjee, and Higgins (2016) estimate welfare effects of patent challenges, finding consumers gain $42 billion from entry associated with challenges, compared to producer losses of $32.5 billion.

patent was wrongly issued (invalid), not infringed by the proposed generic product, or unenforceable. Accordingly, the drug maker applies for approval notwithstanding the unexpired brand patent. In this case, the drug maker's application includes a "Paragraph IV" certification, so-named after the relevant clause of the Hatch-Waxman Act, that explains why the patent does not legally cover the generic product.[7] For a new chemical entity, an abbreviated new drug application containing a Paragraph IV certification—a so-called Paragraph IV challenge—can be filed as early as four years after approval. A generic drug maker that uses a Paragraph IV certification must notify the branded drug maker of its submission.

Step Three: Litigation and automatic stay of FDA approval. A legal battle will often ensue at this stage. Upon notification, the brand company may sue the generic for patent infringement before the generic begins selling its product. If the suit is filed in a timely fashion, FDA approval of the generic firm's abbreviated new drug application is automatically blocked for up to 30 months while the suit is considered by a court.[8] If the generic wins the patent litigation sooner than that, the stay expires early. Note that this delay arises even if the patent is invalid (for example, because the invention is obvious) or not infringed (for example, because the generic drug product uses a different technology), and even if the patent should not have been listed in the Orange Book in the first place (for example, because it does not pertain to the drug in question).

Generic drug makers face a potential collective action problem in this setting: no firm may be willing to pay the costs of litigating the validity of a brand-name patent if the result of a successful patent invalidation by one firm simply opens the door to free entry by all other generic firms, resulting in a crowded market with low margins. For this reason, the Hatch-Waxman Act created a special incentive for generic firms to bear the expense of a Paragraph IV challenge. The first generic drug maker to file such a challenge is eligible for a 180-day exclusivity period to market the generic drug before other generics may enter the market. In this way, the first filer can capture the lion's share of sales and establish a leading market position. Thus, generic exclusivity is a lucrative "bounty" worth hundreds of millions of dollars for a blockbuster drug. Generic exclusivity can also create a bottleneck for subsequent filers as they wait for the period to expire. Once generic entry occurs, the branded firm often raises its price, earning a higher margin on a minority of brand-loyal customers.

[7] Another option, in the case of patent claims on a method of use of the drug, is to avoid the patent by means of a "section viii" statement, which asserts that the claims do not cover any use for which the generic seeks approval.

[8] If the abbreviated new drug application is filed less than five years after approval of a new chemical entity (recall that this can occur as soon as four years after approval), then the expiration date of the stay is extended. Under the extension, the stay expires 7.5 years (five years plus 30 months) after the date of approval of the brand-name drug.

Hatch-Waxman and Entry: Orange Book Patent Listings, Challenges, and Generic Approval

The Growth of Patenting

The pharmaceutical industry has been traditionally classified as a "discrete product" industry, in which one or relatively few patents cover a single product (Levin et al. 1987; Cohen et al. 2000). The classic "discrete product" drug has one strong patent, covering its active ingredient.

But in the decades since Hatch-Waxman, real-world drug patenting has diverged from this initial pattern. Here we extend the analysis in Hemphill and Sampat (2011), which reported growth in Orange Book patents over time. Our starting point, using data from the Food and Drug Administration, is the set of new molecular entities—drugs containing a novel active ingredient—approved between 1985 and 2020 (FDA 2023a). To gain a complete picture of patenting for these drugs requires linking various editions of the Orange Book. (Each edition of the Orange Book contains a snapshot of unexpired patents.) The NBER Orange Book dataset does this, relying upon digitized archival versions of the Orange Book between 1985 and 2009 and digital versions through 2016 (Durvasula et al. 2023). We update the Orange Book data to include all editions to 2023. The online Appendix provides details.

The final dataset includes 826 drugs approved between 1985 and 2020, and their 4,446 patents.[9] Where the same patent protects more than one drug, it is counted more than once; technically, there are 4,446 drug-patent "pairs" (and 4,217 unique patents).

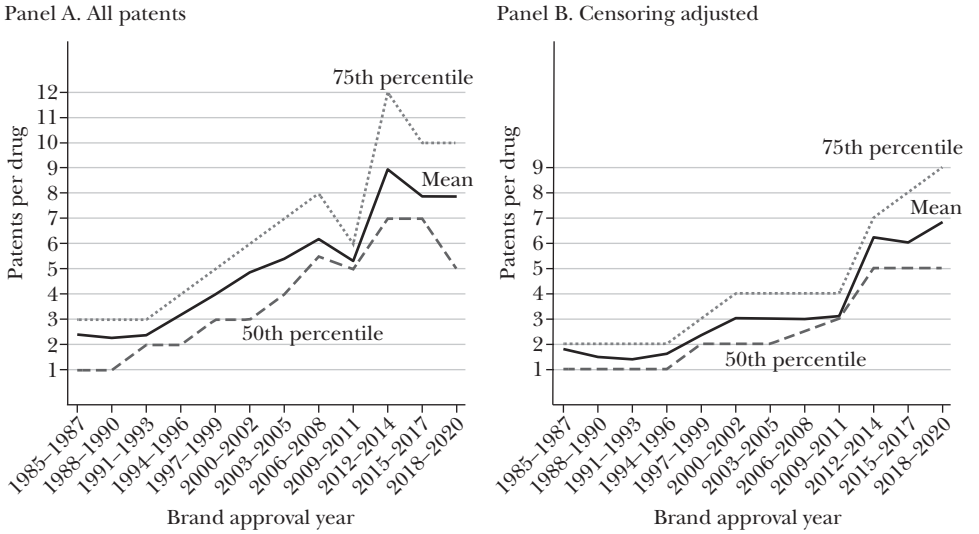
Figure 1 shows the growth in patents over time. For the 1985–1987 approval cohort—that is, branded drugs approved between 1985 and 1987—the median number of patents per drug is one. The median increases over time to seven patents per drug by the 2012–2014 cohort. The mean also increases, from 2.4 patents per drug to more than eight in later years. At the 75th percentile, patenting rises from three patents per drug to ten.

Secondary Patents, Evergreening, and Nominal Patent Term

To understand the increase, it is important to distinguish two types of patent. A "primary" patent covers the active ingredient of a drug. "Secondary" patents cover ancillary aspects of the drug, such as chemical variants, alternative formulations, and methods of use. The growth in median and mean patents for new drugs is driven by an increase in secondary patenting. Within the 1985–1987 cohort, 54 percent have at least one secondary patent based on patent categorizations from IQVIA's Ark Patent Intelligence (see the online Appendix for details). By the mid-2010s, this is true of nearly all drugs, a trend sustained through the 2015–2017 cohort (the most recent for which we have patent categorization data).

[9] In addition to patents, nearly all of these drugs (802/826) receive regulatory protection as new chemical entities.

Figure 1
Patents per Drug, 1985–2020



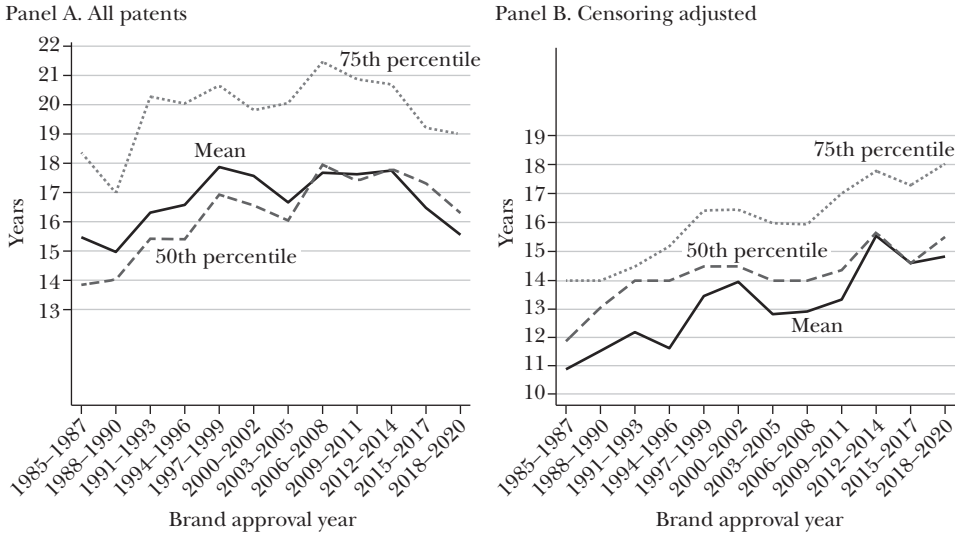
Source: FDA (2023a, b) and NBER (2023). See the online Appendix for details.
Note: Panel A shows the total number of patents per drug for the 826 patented new molecular entities approved between 1985 and 2020. Panel B is restricted to patents listed within three years of drug approval to address right-censoring. Reissued patents are linked to their original patents and counted as a single patent.

Applications for these secondary patents are disproportionately filed later: on average the filing date of primary patents is twelve years before brand approval, compared to seven years before brand approval for secondary patents.[10] Ten percent of drugs have at least one patent filed after drug approval: these are almost all secondary patents. Later applications result in later patent expirations, given that patents expire 20 years after application (or, before 1995, 17 years from issue).[11] Later-expiring secondary patents can thus lengthen (in the language of Hemphill and Sampat 2011) the "nominal" patent term: the time from FDA approval of a drug to the expiration date of its last expiring Orange Book patent.

Figure 2 shows the increase in nominal term over time, from about 15 years (at the mean) in the 1985–1987 cohort, to 18 years in the 2000–2002 cohort. While the data from all Orange Book patents shows a slight decline in the most recent cohorts, the second panel (excluding late listed patents) shows a steady increase over time, suggesting the recent dip is due to censoring.

[10] We use "effective" filing date (also called the priority date) for this calculation, based on data from the Google Patents Public Dataset (2023).
[11] Not all late applications result in late expirations. For a discussion of the complexities, see Lietzan and Acri (2020).

Figure 2
Nominal Patent Term, 1985–2020



Source: FDA (2023a, b) and NBER (2023). See the online Appendix for details.
Note: Panel A shows the nominal patent term (years from brand approval to last expiring patent for the drug) for 826 patented new molecular entities approved between 1985 and 2020. Panel B restricts to patents listed within three years of drug approval to address right-censoring.

This extension of patent term through secondary patents is part of a broader set of entry-delaying efforts called "evergreening" by its critics, and "life-cycle management" by practitioners. Even absent nominal term extension, when both primary and secondary patents are listed on the Orange Book, the result can be a "thicket" of temporally overlapping claims, which a generic drug maker must either address or wait out to market a competing product.

A further reason that secondary patents have been more controversial than pharmaceutical patents in general, along with potentially extending exclusivity, is that they are of lower average quality from a legal perspective, in the sense that they are less likely to meet traditional standards of patentability (novelty and non-obviousness) and thus less likely to be valid. Such patents are sometimes granted erroneously by a US Patent and Trademark Office operating under resource and information constraints (Lemley and Shapiro 2005; Lemley and Sampat 2012; Frakes and Wasserman 2023). In addition, some of the secondary patents listed on the Orange Book may not be infringed by a proposed generic product. As reported above, some secondary patents are filed even after drug approval, and thus less plausibly relevant to the drug as approved.

These patterns suggest a concerning scenario in which secondary patents are used to create expense, uncertainty, and delay for generic drug makers, thus

deterring some entry that would otherwise be attempted. A branded drug maker has an incentive to apply for and to list as many patents as possible, and thereby force a sequence of Paragraph IV challenge, litigation, and 30-month stay. The resulting litigation, aside from delaying entry, also promotes a dynamic in which the generic firm has an incentive to settle the lawsuit with a promise to accept a later entry date in order to preserve its entitlement to the 180-day bounty (Hemphill 2006). The bounty, often a major source of profit for the generic drug maker, is placed at risk if the generic litigates to judgment instead of settling.

Patent Challenges and Effective Market Life

Patent challenges in pharmaceutical markets have grown sharply over time. About 21 percent of drugs in the 1985–1987 approval cohort have one or more patents challenged by generics; this figure rises to 55 percent by 2000–2002. Nearly 80 percent of drugs approved between 2005 and 2010 face challenges, though the share declines in later years.

These challenges can and increasingly do start early in the life of a drug. There is a spike in challenges at four years after approval—for new chemical entities, the earliest point they are allowed. For most commercially important drugs, drug makers not only know to expect a challenge, but also when to expect it.

Some have argued that the rewards from generic challenge are too large: that the high benefit from a successful challenge (particularly the profits during the 180-day bounty period) or profitable settlement, and comparatively low cost, leads to prospecting by generics. Voet (2005) comments that generic drug makers "rely on the law of averages—if you place enough bets, you are sure to win a few of them." Several papers also suggest that challenges disproportionately target patents on blockbuster drugs (Grabowksi and Kyle 2007; Higgins and Graham 2009; Grabowski et al. 2017). This is concerning because, based on commonly cited estimates of the returns to research and development spending, blockbusters have an outsized role in helping branded drug makers cover the average cost of research and development (DiMasi, Grabowski, and Hansen 2016).

Our research on these issues suggests a more nuanced story. While higher-sales drugs indeed are more likely to draw challenges, in part that is because those drugs also have more patents per drug, including a larger number of secondary, late-expiring patents driving a long nominal patent term (Hemphill and Sampat 2011). In previous work examing patents, challenges, and generic entry for drugs with first-time generic entry between 2001 and 2010, we found that even within drugs, secondary patents are disproportionately challenged (Hemphill and Sampat 2012). From this perspective, the Hatch-Waxman regime confers upon generic firms a high-powered incentive to give weak patents on important drugs a strong second look (Engelberg 1999; Bulow 2004).

Generics are also more likely to be successful when challenging secondary patents. Hemphill and Sampat (2013) show that for patent challenges litigated to completion, branded firms win on primary patents 92 percent of the time, but on secondary patents only 32 percent of the time. While litigation outcomes are

notoriously difficult to interpret when settlement is an option, for a smaller subset of cases adjudicated in an era where settlement was less common, the branded firm wins nearly all litigation resulting from challenges to primary patents, and generic challengers typically prevail on secondary patents, consistent with the legal understanding of secondary patents as weaker or lower quality.

Successful challenges on secondary patents (including settlements with an early entry date) can thus offset evergreening. A drug's effective market life—the time from brand approval to first generic entry—is typically less than its nominal patent term. For drugs with first generic approval (specifically, of an application on the same active ingredients) between 2000 and 2023, the average effective market life (measured by the timing of generic approval) is 13.0 years, compared to a nominal patent term of 17.5 years.

Figure 3 illustrates how these dynamics play out over the sales distribution for drugs with first generic approval between 2008 and 2023, using sales data from SSR Health (2024). Patent accumulation and nominal patent term increase with sales. Higher sales drugs also draw more challenges and have a shorter effective market life. Challenges to late-expiring secondary patents may not be the whole story. Hemphill and Sampat (2012) present evidence that for more lucrative drugs, patent challenges to primary patents are also more common. The incentive to challenge primary patents may be buttressed by a racing dynamic. If one generic firm would otherwise wait until the primary patent expires, it may be induced to challenge that patent by its expectation that others will do so, lest the firm miss out on exclusivity (Hemphill and Sampat 2012; Grabowski et al. 2017).

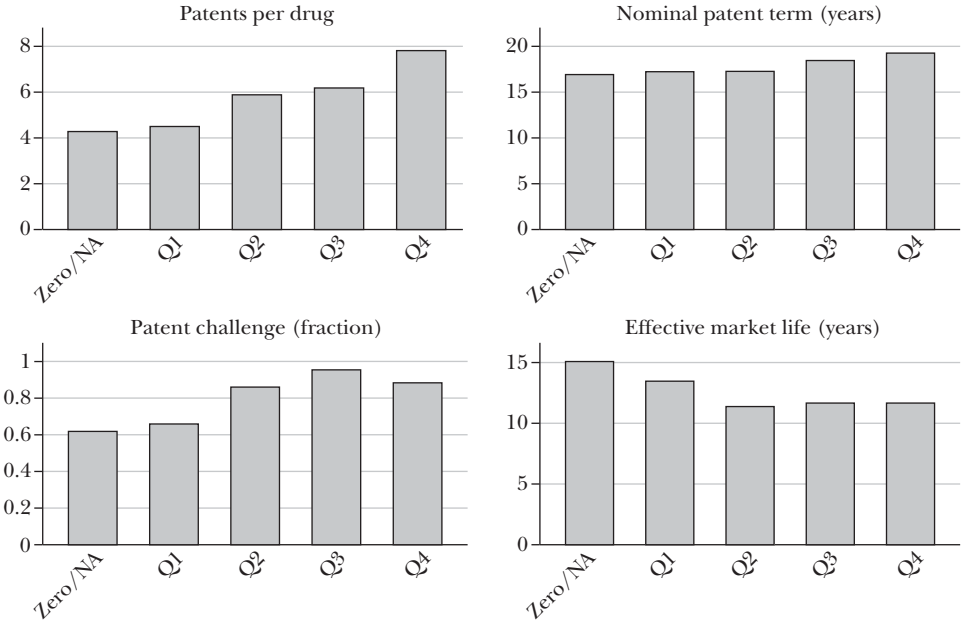
Despite these complex interactions, effective market life has remained fairly stable over time, as shown in Figure 4, generally ranging between 12 and 14 years since 2000, though it dips slightly in the most recent cohort. The general stability of effective market life over time, reported by Hemphill and Sampat (2012), Danzon and Furuakwa (2011), and others, is notable, given the growth of patents per drug, secondary patenting, and nominal term in the four decades since Hatch-Waxman.

As further evidence on the nature of patents that are most relevant, we examine the share of patents that potentially matter for the timing of generic approval. We classify a patent as "non-binding" for a drug if the first generic approval occurs before the expiration of the patent. For drugs with generic approval between 2000 and 2023, we calculate the share of patents that are non-binding for generic entry. We find that 65 percent of secondary patents are non-binding, compared to 24 percent of primary patents. This result is consistent with our previous work, and that of other scholars, showing the relative weakness of secondary patents.

Paying for Delay, and Other Games

One way for a branded drug maker to hinder generic entry, and thereby reap the benefits of delay discussed above, is to pay the would-be generic entrant to postpone or abandon its efforts to enter the market. This strategy, a form of collusion, exploits the fact that monopoly profits are typically larger than total profits in a competitive duopoly (Gilbert and Newbery 1982). This strategy predates the modern

Figure 3
Patents, Challenges, and Effective Market Life, by Sales Category

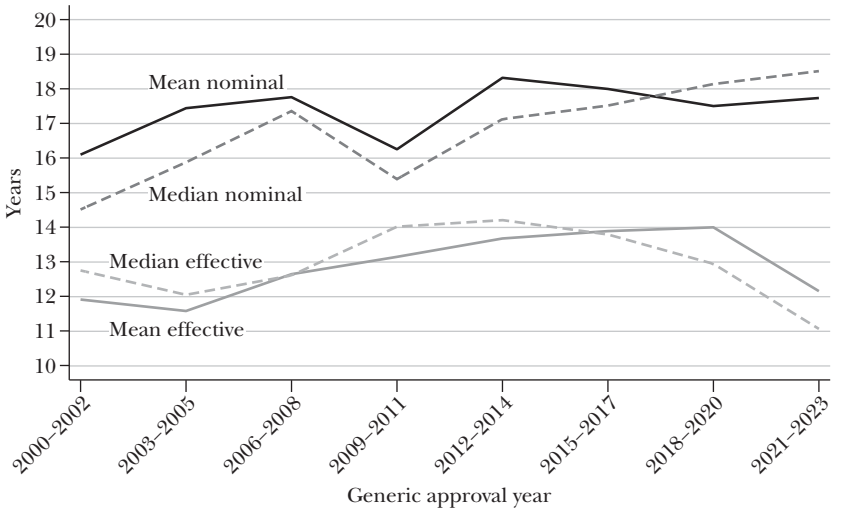


Source: FDA (2023a, b, c, d), NBER (2023), and SSR Health (2024). See the online Appendix for details.
Note: This figure shows patents per drug, nominal patent term (time from brand approval to last expiring patent for the drug), whether there was a Paragraph IV patent challenge, and effective market life (time from brand approval to first generic approval for the drug's active ingredient(s)), by sales category, for the 292 drugs with first generic approval between 2008 and 2023. Sales categories are based on SSR Health data from the calendar year prior to first generic approval. All sales are adjusted to 2022 dollars. Drugs are grouped into quartiles of non-zero sales, where Q1 represents the lowest quartile (mean sales of $118 million) and Q4 represents the highest quartile (mean sales of $2.9 billion). The 118 drugs without SSR Health data in the year before generic entry are plotted separately, as they may represent either zero-sales drugs or those outside SSR Health's sampling frame.

pharmaceutical industry. In the nineteenth century, Bell Telephone filed a patent suit against Western Union, a competitor in telephony. Under their 1879 settlement, Bell paid Western Union many millions of dollars and Western Union agreed to stay out of the telephone business for more than a decade (Farrell and Chicu 2018, citing Brock 1981). Branded drug makers use the same playbook, with the branded firm in the position of Bell, and a generic entrant playing the role of Western Union. For the monopolist, payment makes sense as a way to avoid duopoly. As for the entrant, the incentive to compete aggressively is neutralized by the payment.

In the pharmaceutical context, the payment is made as part of a settlement of Paragraph IV patent litigation. The patentee pays an alleged infringer to abandon its bid for competitive (and allegedly infringing) entry. Paying off one generic drug maker might seem pointless, given the risk that others would enter anyway

Figure 4
Nominal Patent Term and Effective Market Life, 2000–2023



Source: FDA (2023a, b, c, d) and NBER (2023). See the online Appendix for details.
Note: This figure shows trends by generic approval year in nominal patent term (time from brand approval to last expiring patent for the drug) and effective market life (time from brand approval to first generic approval for the drug's active ingredient), for the 382 drugs with first generic approval between 2000 and 2023.

or demand payments of their own. But as discussed earlier, "first filers" under the Hatch-Waxman regime are potentially eligible for 180 days of generic exclusivity, before other generics can come in. Moreover, later challengers are impeded due to a regulatory bottleneck that inhibits approval of later filers (Hemphill and Lemley 2011). Thus, later filers have a relatively low incentive to pursue early entry.

In an ordinary patent settlement, one would expect payment to flow from infringer to patentee, as compensation for expected damages from past infringement. Thus, an observed payment from the patentee to the alleged infringer is a notable anomaly. For this reason, pay-for-delay deals are sometimes called "reverse payment" settlements. An ordinary settlement reflects the strength of the underlying patent—that is, whatever limitation on entry the patentee could secure based on its probability of winning the infringement suit. By contrast, when the patentee instead makes an additional payment to sweeten the deal, the result is less competition than the patentee could expect by asserting the patent alone. The payment, in other words, is for additional generic delay, compared to what is legitimately achieved by the patent alone.

Accordingly, an observed and otherwise unexplained payment is a basis for inferring that the branded firm is profiting at the expense of consumers (Shapiro 2003; Hemphill 2006; Edlin et al. 2013). This inference has been adopted by the US Supreme Court in an important antitrust case addressing reverse payment

settlements (*FTC v. Actavis, Inc.*, 570 US 136 [2013]). In the wake of this decision, lower courts have struggled with what (beyond cash) counts as a payment, such as overpayment for services provided by the generic firm, forgiveness of a debt owed by the generic (such as liability for patent infringement on other drugs), or early generic entry on other drugs or in other jurisdictions (Hemphill 2009).[12] Lower courts have also considered what (if any) role is played by the merits of the underlying patent litigation that gave rise to the settlement.

Branded drug makers use other strategies to delay generic entry. One such strategy is to file a large set of "citizen petitions" shortly before generic approval, to which the Food and Drug Administration is legally required to respond, thereby delaying generic approval (Carrier and Minniti 2016). The stability of effective market life as measured by the timing of approval, shown in Figure 4, suggests that the aggregate effects of these tactics may be limited.[13] But for blockbuster drugs, even a few months of additional exclusivity can make them worthwhile strategies for branded drug makers to pursue.

A further strategy called "product hopping" increases the impact of delayed entry (Carrier and Shadowen 2017). A branded firm can shift patients and doctors from a drug facing imminent generic competition to a new version with stronger or longer-lived exclusivity. If the "product hop" is accomplished before the generic version of the old drug is approved, there is no foothold for generic substitution to take place with a pre-existing base of customers.

Hatch-Waxman and Innovation: Evaluation of Patent Term Restoration

The other side of the Hatch-Waxman compromise is patent term restoration—that is, provisions to add back time lost to clinical trials and regulatory review, which drug companies and some academics argue have been eroding patent terms and thus blunting innovation incentives.

Of the drugs in our dataset, 69 percent (571/826) have a patent extended under these provisions. The average extension is about three years (1076 days at the mean, 1022 at the median). Notably, for 30 percent (169/571) of the extensions, the five-year maximum extension (or two years for certain drugs) is binding.[14] For

[12] One common form of compensation involves the brand's ability to compete with the generic entrant by deploying an "authorized generic" marketed under the brand's original drug approval. By agreeing not to launch an authorized generic, the branded firm confers a benefit on the generic (at its own expense), a sacrifice that can be used to secure delayed entry by the generic.

[13] One caution is that we are measuring generic entry by approval, not actual generic launch. The strategies discussed above can delay entry past the point of generic approval, at least in some cases. Understanding the extent of this dynamic is an important topic for future research. See the online Appendix for more details on the distinction between approval and launch, and potential empirical approaches to measuring the latter at scale.

[14] By our count, 83 extensions were capped at five years. For another 86 "pipeline" drugs (see footnote 4 above for an explanation), the extension was capped at two years.

another 26 percent, the post-approval cap of 14 years binds. Thus, of the drugs that receive patent extensions at all, only a minority get the full benefit of the restoration formula, a striking demonstration of the incompleteness of term restoration. Even for non-capped drugs, only half of the time lost in clinical trials is recouped (plus all of the time spent in regulatory review of the new drug application).[15]

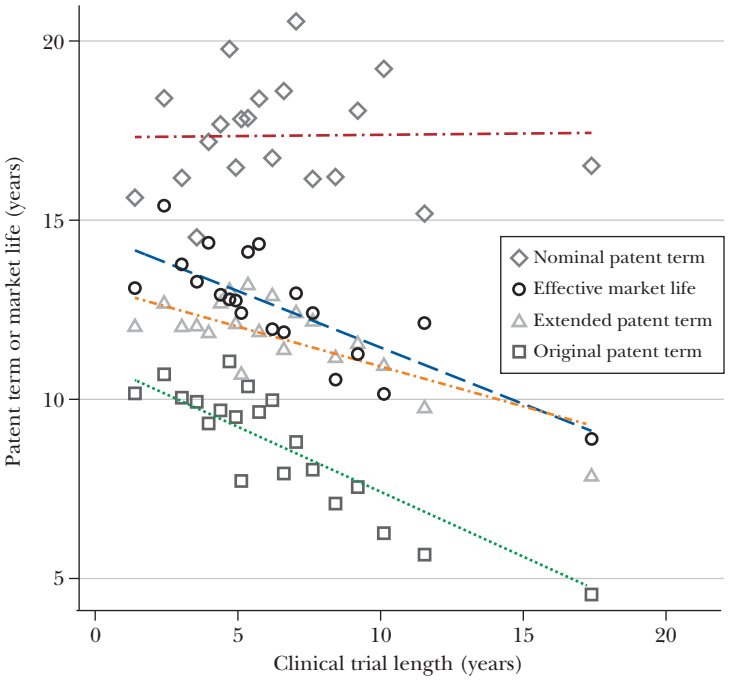
In the previous section, we emphasized that new drugs increasingly have numerous patents. In choosing which patent to extend, the drug maker faces a tradeoff (Eisenberg 2012). Secondary patents often expire later but, as discussed above, are more vulnerable to challenge. For drugs approved since 1985 with an extended patent, primary patents account for 73 percent of the extensions, compared to just 28 percent of all patents on these drugs. This provides a "revealed preference" measure of which patents firms view as being most important for their drugs (Hemphill and Sampat 2011).

Lietzan and Acri (2020) show that patents with longer trials have shorter terms, even after extension. We update and extend their analysis for our sample of drugs. Figure 5 shows the relationship between clinical trial length and different measures of patent term or market exclusivity for drugs with at least one extended patent and generic entry between 2000 and 2023. There is a strong negative relationship between trial length and original years of patent term, before patent term restoration. The term extension shifts the level of term upwards and reduces its slope (in expected value), but it is still negative. This result is unsurprising given the term extension formula and its built-in caps. Nominal patent term, sustained by all patents on the drug (not just the extended one) is flat: additional patents break the relationship between clinical trial length and patent term. However, as we have argued above, nominal patent term does not tell the full story, because secondary patents are often successfully challenged, especially for important drugs. The relationship for effective patent life, the time to actual generic entry, and clinical trial length is once again negative.

Bringing the facts together: Hatch-Waxman term extensions restore life lost to clinical trials, but do not fully reverse the negative relationship between clinical trial length and the term of the extended patent. Secondary patents on the same drugs appear to undo the relationship at the drug level. However, these secondary patents are challenged by generic firms through the Paragraph IV process, leaving only the extended (typically primary) patent standing, resulting in a negative relationship between trial length and effective market life. These results are consistent with the conclusion of Lietzan and Acri (2020) that incomplete restoration is likely to distort drug development away from the most difficult challenges, such as Alzheimer's disease and pancreatic cancer.

[15] These calculations are based on a compilation of patent extensions maintained by the Patent Office. While this compilation is generally accurate, it does not fully reflect certain changes made to the (pre-extension) patent expiration date after an initial extension decision. Earlier editions of the compilation attempt to make relevant adjustments. In the online Appendix, we show that the results are robust to using this alternative data source.

Figure 5
Patent Term and Effective Market Life versus Clinical Trial Length



Source: FDA (2023a, b, c, d), NBER (2023), USPTO (2023a, b, c), Hemphill and Sampat (2025b, c, d), and NARA (2023). See the online Appendix for details.
Note: This chart shows binned scatterplots (20 bins each) of nominal patent term (time from brand approval to last expiring patent for the drug), effective market life (time from brand approval to first generic approval for the drug's active ingredient), and original and extended term of the patent with extension, plotted against clinical trial length, for 382 drugs with an extended patent and first generic approval between 2000 and 2023. Dashed lines show linear regression fits through the underlying data.

Whether the resulting level of effective market life is too high or too low is not obvious, but the distortion is troubling. Expectations about effective market life (including assessment of strength of different patents) shape the choices that firms make about research and development. Shorter exclusivity periods for longer development times may lead firms to reduce investments in long-term research. Consistent with this idea, Budish, Roin, and Williams (2015) show that firms' investments in cancer research are distorted away from drugs that require long clinical trials.

Proposals for Hatch-Waxman Reform

The 1984 Hatch-Waxman Act uses a combination of patents and regulation in seeking to promote both dynamic and static efficiency in pharmaceuticals. Based on

40 years of experience under this regime, how successful has it been? In this section, we take stock and set out some ideas about improvements to the patent listing, challenge, and extension provisions.

Competition

By creating a pathway to generic entry, Hatch-Waxman helped to fuel the growth of the generic drug industry. Yet our discussion also suggests problematic manipulation of the regime. Brand-name drug producers have a strong incentive to obtain and list secondary patents, even those of doubtful validity or relevance. The extra patents drive up the nominal patent term and potentially raise barriers to entry for generic firms. That said, the data suggest that patent challenges, incentivized by the 180-day exclusivity bounty for some generic entrants, seem to be effective in targeting and largely negating these patents. Settlements and other gaming interfere with the machinery in important specific cases, but overall the challenge process does appear to ratchet back the nominal term.

But might other approaches to the same end be possible? After all, our current system of patent accumulation, listing, challenge, and costly litigation both wastes resources and results in a selective lengthening for some drugs but not others, with a final exclusivity period that is unlikely to correspond to social value.

One avenue for reform is to do a better job in preventing secondary patents from being issued in the first place. These proposals are sometimes framed as responses to the proposition offered by Lemley (2001) that it might be "rationally ignorant" for the Patent Office to screen lightly in general, given that most patents do not matter, relying on litigation to invalidate patents that turn out to be important. Rather than rely on litigation to challenge patents after issuance, there could be greater up-front investment in screening by the Patent Office. Increasing the review time of pharmaceutical patents by the examiners can be highly cost-effective, through reducing litigation costs and speeding generic entry (Frakes and Wasserman 2023).

Perhaps even more promising, pharmaceuticals offer a rare context in which the knowledge of which patents actually matter for competition is publicly revealed, thanks to the Orange Book listing, which narrows the inquiry (Hemphill and Sampat 2012). Even if it is a practical impossibility to do an intensive review of all of the more than 600,000 patent applications filed at the Patent Office each year, selective review of pharmaceutical patents does seem feasible. In the 2023 edition of the Orange Book, there are only 5,633 unexpired pharmaceutical patents overall, of which only 645 were newly added since 2022.

To ensure validity of these listings, the Patent Office could provide an intensive second layer of review to pharmaceutical patents at the time of Orange Book listing (Hemphill and Sampat 2011, 2022). For drug makers, one benefit of this approach is that it could effectively "gold plate" high-quality patents against further litigation (Lemley, Lichtman, and Sampat 2005), limiting the "prospecting" type challenges that worry some observers (Higgins and Graham 2009).

Another approach is to alter the process for clearing out invalid patents (Hemphill and Sampat 2022). In 2011, Congress established the Patent Trial and Appeal

Board to adjudicate patent validity at a lower cost compared to traditional litigation. However, this process is not well integrated with the Hatch-Waxman framework, limiting the impact of its procedures for pharmaceuticals (Rai et al. 2022). The main issue is that if a generic drug maker goes before the Board and wins, it does not secure the 180-day exclusivity. Changing the Hatch-Waxman Act to incorporate this new mechanism for patent challenges is another important potential reform.

Validity of patents is one issue. Whether a patent actually covers the drug is another. As noted above, currently the Food and Drug Administration performs no review of whether a listed patent actually pertains to the drug. One novel response was undertaken by the Federal Trade Commission (2023, 2024), which challenged as improper the listing of a large set of patents. Going forward, the regulator should find a way to move beyond a purely ministerial role in administering patent listings (Eisenberg and Crane 2015; Hemphill and Sampat 2022; Contreras and Rai 2023). As Contreras and Rai (2023) note, private patent pools scrutinize each patent for essentiality to the standard, suggesting the feasibility of a less searching scrutiny of each Orange Book patent for compliance with regulatory requirements. At a minimum, the agency should require more precise specification of what exactly the listed patents cover and deny listing of those that do not apply to the approved product (Hemphill and Sampat 2023).

A final avenue for reform would directly alter the branded firm's cost-benefit analysis in erroneously asserting a patent that is invalid or not infringed by the proposed generic product. Currently, a branded drug maker experiences no adverse consequences when it blocks generic entry by means of the automatic stay, using a patent that is later shown to be invalid or not infringed by the generic drug product. Imposing a financial obligation on the branded firm in these circumstances would cause it to internalize some of the social costs imposed by its erroneous assertion.[16]

The primary objective of current proposals to restrict secondary patents is to reduce drug expenditures. How big an impact might these kinds of changes have? To examine this, we focused on the 399 drugs from our dataset approved between 1985 and 2019 that had at least one unexpired patent listed in the 2019 Orange Book. Overall, these drugs accounted for $134 billion in sales in 2019, among drugs in SSR Health's drug sales database (SSR Health 2024). Of these drugs, 6 percent had a primary patent only in 2019, 40 percent a secondary patent only, and 54 percent had both types. The finding that nearly all drugs from our sample that are on-patent in 2019 have a secondary patent is consistent with our data above on the growing prevalence of secondary patenting, and consistent with the idea that secondary patents are protecting economically valuable drugs.

[16] Other jurisdictions provide instructive analogies. For example, Australian law provides for the branded firm to make compensatory payments to the state and to competitors if it secures a temporary legal block of entry while lacking a reasonable basis for doing so (Therapeutic Goods Act of 1989, § 26D). Similarly, a 2021 judicial opinion of Israel's Supreme Court prescribed disgorgement, payable to the generic firm, of improperly acquired profits secured due to an improper delay of generic entry (*Unipharm v. Sanofi*, CivA 2167/16 [2021]).

But the majority of these on-patent drugs—60 percent—are protected in whole or part by primary patents. These drugs are associated with an even larger fraction of SSR Health-recorded sales: 76 percent. Through eliminating or curtailing secondary patents could meaningfully reduce drug expenditures, the incremental value of such proposals is limited by the continued relevance of strong primary patent protection. To look at this another way, patent challenges have already had an impact that these figures do not capture, because many important drugs have already gone generic by 2019 thanks to challenges on secondary patents.

Innovation

While curtailing secondary patents (through challenges or additional reforms like those above) can hasten generic entry, doing so comes at the cost of the exclusivity period enjoyed by the innovator. Given the heavy reliance of drug companies on patents for recouping their substantial research and development investments, this raises a concern about innovation incentives. For example, Higgins and Graham (2009) argue that patent challenges might be "tipping the balance" of Hatch-Waxman towards generics and away from innovators, sacrificing dynamic gains for static benefits. The strength of this argument depends on whether optimal patent term is closer to nominal patent term, including that sustained by secondary patents, or patent term sustained by primary patents alone. We do not know the magic number for an optimal fixed patent term. But whatever it may be, we are skeptical that the right way to get there is through accumulation of patents of questionable relevance or validity.

We can say with more certainty that the term restoration provisions of the act need rethinking. The partial restoration of time lost in trials, including the five-year cap, means that drugs with longer development time get less time on the extended patent, which we have shown translates to a shorter period of effective market exclusivity at the drug level. This penalizes research with long development times. If the goal is uniformity, then complete restoration, or simply beginning patent term at approval, would be more appropriate. A version of this approach that may be palatable to branded and generic firms alike would be a one-and-done approach: allowing firms to list one patent, with time lost to trials completely restored, but only one.

However, uniform exclusivity for all drugs may not be the optimal policy (Budish, Roin, and Williams 2015). An alternative is to index the reward of exclusivity to the social benefits of research, or to time-to-market (Roin 2013). International agreements around patent harmonization limit the feasibility of implementing nonuniform patent terms in general. But in pharmaceuticals, one can imagine how the Food and Drug Administration might administer new exclusivity modeled on the Hatch-Waxman approach for extending patent terms. This approach might serve as an additional innovation policy lever (on top of or even instead of patents), which might be more proactively deployed to shape both the rate and direction of pharmaceutical innovation in a more flexible and nuanced way than possible through patent policy alone (Eisenberg 2012). A system based mainly on nonpatent exclusivities may also avoid the costs of the Hatch-Waxman Act's approach to balancing dynamic and static efficiency—which relies on the complex interplay of

patent restoration, listing, and litigation to determine effective exclusivity terms—potentially reducing uncertainty for branded firms and generics alike.

Conclusion

In the 40 years since Hatch-Waxman, a pattern has taken hold in pharmaceutical markets. Branded drug makers accumulate patents on drugs and list them on the Orange Book, including a primary patent on the active ingredient (typically the one that is chosen for extension) and additional peripheral secondary patents. Patent challenges can commence four years after the approval of a new chemical entity—or sooner for other drugs—triggering litigation and sometimes settlement. Patent accumulation and speed and intensity of challenges are greater for blockbuster drugs. At the end of the process, much of any extra nominal term generated beyond the primary patent is ratcheted back. Primary patents sustain most of the effective market life of drugs, which has generally oscillated between 12 and 14 years since Hatch-Waxman.

The right amount of exclusivity to balance innovation and access was unknown in 1984 when Hatch-Waxman was enacted, and in our view it remains unclear today. Various idiosyncratic features of the process we have described, including of the term restoration process, might be exploited to collect better evidence on the causal impact of marginal changes to patent term on innovation and competition. At the same time, the ritualistic aspects of the "listing, challenge, litigation" process caution against using patent challenges or their outcomes as exogenous shocks to patents or patent term: these steps are largely anticipated by all parties involved.

The administrative data created by Hatch-Waxman are a promising source for economic research. Orange Book patent listings provide information linking patents to their products—which is not systematically possible in other industries—potentially a boon for empirical research on innovation (Durvasula et al. 2023). The major caveat is that all Orange Book patents are not created equal, and for many research questions primary and secondary patents should be considered separately. Relatedly, both researchers and policymakers describing the growth of patenting and patent terms need to distinguish between nominal patent term and effective market life: just because a patent is listed on the Orange Book does not mean it is likely to impede generic entry.

Throughout this essay, we have focused on drugs that are new molecular entities, the canonical Hatch-Waxman drugs around which the regime was designed. Some of the dynamics may be different for drugs that are line extensions, not new molecules but modifications to earlier drugs. The challenge and generic entry process may also be different for drugs that are more difficult to imitate, or where FDA regulations for entry are more cumbersome (on the breakdown of the Hatch-Waxman system for drug-device combinations, see Feldman et al. 2022; Reddy et al. 2023).

Moreover, a large share of US drug expenditures is now for biologic drugs, as opposed to the "small molecule" drugs created by chemical synthesis that are

governed by Hatch-Waxman. Biologic drugs are not subject to Hatch-Waxman's patent listing and challenge regime, nor its mechanism for generic entry. But even here, there are some similarities and some potential lessons from Hatch-Waxman. For biologics, as for other drugs, a major policy challenge remains implementing a system of listing patents that provides clarity to potential entrants, while at the same time ensuring that invalid or improperly listed patents do not impede generic entry.

Much of our discussion has considered variations on the Hatch-Waxman approach, which presumes that limited patent term exclusivity is the best way to incentivize innovation, and generic competition the best way to promote lower prices and access. But this dichotomy may blind us to possibilities for more fundamental policy change. In the 40 years leading up to Hatch-Waxman, policymakers employed mechanisms for promoting innovation without sacrificing access, including direct government funding of applied research and active use of government procurement as an innovation incentive (important in World War II, but largely absent afterwards until the Covid-19 vaccine development effort). Prize-based mechanisms and formal advance market commitments are other options. There are also approaches beyond Hatch-Waxman to implement the tradeoff between innovation and access, such as compulsory licensing or conditioning legal protection on therapeutic efficacy (two features, abandoned along the way, of the original legislation that resulted in the 1962 Kefauver Act). The 40-year milestone also presents an opportunity to revisit these and other policy levers—beyond patent extensions and generic entry—to promote the dual goals of dynamic and static efficiency in pharmaceuticals.

■ *Sampat's research was supported by a grant from the National Institute of Healthcare Management for a project titled "Can Improving Pharmaceutical Patent Quality Promote Competition and Reduce Drug Prices?" We thank Daniel Francis, Margaret Kyle, Lisa Larrimore Ouellette, Rachel Sachs, Michal Shur-Ofry, and workshop participants at the Brookings Institution, New York University, and the University of Pennsylvania for helpful comments, and Erika Lietzan for suggestions and data related to the patent term extension analyses.*

References

AAM (Association for Accessible Medicines). 2024. *The US Generic and Biosimilar Medicines Savings Report.*

Branstetter, Lee, Chirantan Chatterjee, and Matthew J. Higgins. 2016. "Regulation and Welfare: Evidence from Paragraph IV Generic Entry in the Pharmaceutical Industry." *RAND Journal of Economics* 47 (4): 857–90.

Brock, Gerald W. 1981. *The Telecommunications Industry: The Dynamics of Market Structure.* Harvard University Press.

Budish, Eric, Benjamin N. Roin, and Heidi Williams. 2015. "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials." *American Economic Review* 105 (7): 2044–85.

Bulow, Jeremy. 2004. "The Gaming of Pharmaceutical Patents." *Innovation Policy and the Economy* 4: 145–87.

Buttorff, Christine, Yifan Xu, and Geoffrey Joyce. 2020. "Variation in Generic Dispensing Rates in Medicare Part D." *American Journal of Managed Care* 26 (11): e355–61.

Buxbaum, Jason D., Michael E. Chernew, A. Mark Fendrick, and David M. Cutler. 2020. "Contributions of Public Health, Pharmaceuticals, and Other Medical Care to US Life Expectancy Changes, 1990–2015." *Health Affairs* 39 (9): 1546–56.

Carrier, Michael A., and Carl Minniti. 2016. "Citizen Petitions: Long, Late-Filed, and At-Last Denied." *American University Law Review* 66 (2): 305–52.

Carrier, Michael A., and Steve Shadowen. 2017. "Pharmaceutical Product Hopping: A Proposed Framework for Antitrust Analysis." Health Affairs Forefront (blog), June 1. https://www.healthaffairs.org/do/10.1377/forefront.20170601.060360.

Cockburn, Iain, Rebecca Henderson, Luigi Orsenigo, and Gary Pisano. 1999. "Pharmaceuticals and Biotechnology." In *US Industry in 2000: Studies in Competitive Performance*, edited by David C. Mowery, 363–98. National Academies Press.

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh. 2000. "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not)." NBER Working Paper 7552.

Conrad, Ryan, and Randall Lutter. 2019. *Generic Competition and Drug Prices: New Evidence Linking Greater Generic Competition and Lower Generic Drug Prices*. US Food and Drug Administration.

Contreras, Jorge L., and Arti K. Rai. 2023. "Orange Book Over-Declaration of Pharmaceutical Patents: The Advantages of Ex Ante over Ex Post Review." Health Affairs Forefront (blog), December 13. https://doi.org/10.1377/forefront.20231211.190823.

Cutler, David, Angus Deaton, and Adriana Lleras-Muney. 2006. "The Determinants of Mortality." *Journal of Economic Perspectives* 20 (3): 97–120.

Danzon, Patricia M., and Michael F. Furukawa. 2011. "Cross-National Evidence on Generic Pharmaceuticals: Pharmacy vs. Physician-Driven Markets." NBER Working Paper 17226.

DiMasi, Joseph A., Henry G. Grabowski, and Ronald W. Hansen. 2016. "Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs." *Journal of Health Economics* 47: 20–33.

Durvasula, Maya, C. Scott Hemphill, Lisa Larrimore Ouellette, Bhaven Sampat, and Heidi L. Williams. 2023. "The NBER Orange Book Dataset: A User's Guide." *Research Policy* 52 (7): 104791.

Edlin, Aaron, Scott Hemphill, Herbert Hovenkamp, and Carl Shapiro. 2013. "Activating Actavis." *Antitrust* 28 (1): 16–23.

Eisenberg, Rebecca S. 2012. "Patents and Regulatory Exclusivity." In *The Oxford Handbook of the Economics of the Biopharmaceutical Industry*, edited by Patricia M. Danzon and Sean Nicholson, 167–98. Oxford University Press.

Eisenberg, Rebecca S., and Daniel A. Crane. 2015. "Patent Punting: How FDA and Antitrust Courts Undermine the Hatch-Waxman Act to Avoid Dealing with Patents." *Michigan Telecommunications and Technology Law Review* 21 (2): 197–262.

Engelberg, Alfred B. 1999. "Special Patent Provisions for Pharmaceuticals: Have They Outlived Their Usefulness?" *Idea* 39 (3): 389.

Farrell, Joseph, and Mark Chicu. 2018. "Pharmaceutical Patents and Pay-for-Delay: Actavis." In *The Antitrust Revolution: Economics, Competition, and Policy*, 7th ed., edited by John E. Kwoka Jr. and Lawrence J. White. Oxford University Press.

FDA (US Food and Drug Administration). 2023a. *Compilation of CDER New Molecular Entity Drug and New Biologic Approvals*. https://www.fda.gov/drugs/drug-approvals-and-databases/compilation-cder-new-molecular-entity-nme-drug-and-new-biologic-approvals (accessed September 1, 2023).

FDA. 2023b. *Approved Drug Products with Therapeutic Equivalence Evaluations (Orange Book)*. https://www.fda.gov/drugs/drug-approvals-and-databases/approved-drug-products-therapeutic-equivalence-evaluations-orange-book (accessed September 1, 2023).

FDA. 2023c. *Patent Certifications and Suitability Petitions*. https://www.fda.gov/drugs/abbreviated-new-drug-application-anda/patent-certifications-and-suitability-petitions (accessed September 1, 2023).

FDA. 2023d. *Drugs@FDA: FDA-Approved Drugs*. https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm (accessed September 1, 2023).

Feldman, William B., Doni Bloomfield, Reed F. Beall, and Aaron S. Kesselheim. 2022. "Patents and Regulatory Exclusivities on Inhalers for Asthma and COPD, 1986–2020." *Health Affairs* 41 (6): 787–96.

Frakes, Michael D., and Melissa F. Wasserman. 2023. "Investing in Ex Ante Regulation: Evidence from

Pharmaceutical Patent Examination." *American Economic Journal: Economic Policy* 15 (3): 151–83.

FTC (Federal Trade Commission). 2023. "FTC Challenges More Than 100 Patents as Improperly Listed in the FDA's Orange Book." Federal Trade Commission (press release), November 7. https://www.ftc.gov/news-events/news/press-releases/2023/11/ftc-challenges-more-100-patents-improperly-listed-fdas-orange-book.

FTC. 2024. "FTC Expands Patent Listing Challenges, Targeting More Than 300 Junk Listings for Diabetes, Weight Loss, Asthma and COPD Drugs." Federal Trade Commission (press release), April 30. https://www.ftc.gov/news-events/news/press-releases/2024/04/ftc-expands-patent-listing-challenges-targeting-more-300-junk-listings-diabetes-weight-loss-asthma.

Gatwood, Justin, Teresa B. Gibson, Michael E. Chernew, Amanda M. Farr, Emily Vogtmann, and A. Mark Fendrick. 2014. "Price Elasticity and Medication Use: Cost Sharing across Multiple Clinical Conditions." *Journal of Managed Care Pharmacy* 20 (11): 1102–07.

Gilbert, Richard J., and David M. G. Newbery. 1982. "Preemptive Patenting and the Persistence of Monopoly." *American Economic Review* 72 (3): 514–26.

Google Patents Public Data. 2023. *Google Patents Public Datasets.* https://console.cloud.google.com/marketplace/product/google_patents_public_datasets/google-patents-public-data. Accessed September 1, 2023.

Grabowski, Henry, Carlos Brain, Anna Taub, and Rahul Guha. 2017. "Pharmaceutical Patent Challenges: Company Strategies and Litigation Outcomes." *American Journal of Health Economics* 3 (1): 33–59.

Grabowski, Henry G., and Margaret Kyle. 2007. "Generic Competition and Market Exclusivity Periods in Pharmaceuticals." *Managerial and Decision Economics* 28 (4–5): 491–502.

Greene, Jeremy A. 2014. *Generic: The Unbranding of Modern Medicine.* Johns Hopkins University Press.

Greene, Jeremy A., and Scott H. Podolsky. 2012. "Reform, Regulation, and Pharmaceuticals—The Kefauver–Harris Amendments at 50." *New England Journal of Medicine* 367 (16): 1481–83.

Gross, Daniel P., and Bhaven N. Sampat. 2023. "The World War II Crisis Innovation Model: What Was It, and Where Does It Apply?" *Research Policy* 52 (9): 104845.

Hemphill, C. Scott. 2006. "Paying for Delay: Pharmaceutical Patent Settlement as a Regulatory Design Problem." *New York University Law Review* 81 (5): 1553–1623.

Hemphill, C. Scott. 2009. "An Aggregate Approach to Antitrust: Using New Data and Rulemaking to Preserve Drug Competition." *Columbia Law Review* 109 (4): 629–88.

Hemphill, C. Scott, and Mark A. Lemley. 2011. "Earning Exclusivity: Generic Drug Incentives and the Hatch-Waxman Act." *Antitrust Law Journal* 77 (3): 947–89.

Hemphill, C. Scott, and Bhaven N. Sampat. 2011. "When Do Generics Challenge Drug Patents?" *Journal of Empirical Legal Studies* 8 (4): 613–49.

Hemphill, C. Scott, and Bhaven N. Sampat. 2012. "Evergreening, Patent Challenges, and Effective Market Life in Pharmaceuticals." *Journal of Health Economics* 31 (2): 327–39.

Hemphill, C. Scott, and Bhaven Sampat. 2013. "Drug Patents at the Supreme Court." *Science* 339 (6126): 1386–87.

Hemphill, C. Scott, and Bhaven N. Sampat. 2022. "Fixing the FDA's Orange Book." *Health Affairs* 41 (6): 797–800.

Hemphill, C. Scott, and Bhaven N. Sampat. 2025a. *Data and Code for: "Patents, Innovation, and Competition in Pharmaceuticals: The Hatch-Waxman Act After 40 Years."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E2184437V1.

Hemphill, C. Scott, and Bhaven N. Sampat. 2025b. *IND Start Dates Transcribed from Notices of Final Determination (1985–2023).* openICPSR. Version 1. https://doi.org/10.3886/E218443.v1.

Hemphill, C. Scott, and Bhaven N. Sampat. 2025c. *Additional IND Start Dates Transcribed from the Federal Register (1985–2023).* openICPSR. Version 1. https://doi.org/10.3886/E218443.v1.

Hemphill, C. Scott, and Bhaven N. Sampat. 2025d. *Hand-Collected USPTO Patent Reissue Information.* openICPSR. Version 1. https://doi.org/10.3886/E218443.v1.

Higgins, Matthew J., and Stuart J. H. Graham. 2009. "Balancing Innovation and Access: Patent Challenges Tip the Scales." *Science* 326 (5951): 370–71.

IQVIA. n.d. "ARK Patent Intelligence." https://www.iqvia.com/solutions/industry-segments/generics/ark-patent-intelligence (accessed November 2021).

Lemley, Mark A. 2001. "Rational Ignorance at the Patent Office." *Northwestern University Law Review* 95 (4): 1495–1532.

Lemley, Mark A., Douglas Lichtman, and Bhaven N. Sampat. 2005. "What to Do about Bad Patents."

Regulation 28 (4): 10–13.

Lemley, Mark A., and Bhaven Sampat. 2012. "Examiner Characteristics and Patent Office Outcomes." *Review of Economics and Statistics* 94 (3): 817–27.

Lemley, Mark A., and Carl Shapiro. 2005. "Probabilistic Patents." *Journal of Economic Perspectives* 19 (2): 75–98.

Levin, Richard C., Alvin K. Klevorick, Richard R. Nelson, and Sidney G. Winter. 1987. "Appropriating the Returns from Industrial Research and Development." *Brookings Papers on Economic Activity* 18 (3): 783–831.

Lichtenberg, Frank R. 2022. "The Effect of Pharmaceutical Innovation on Longevity: Evidence from the U.S. and 26 High-Income Countries." *Economics and Human Biology* 46: 101124.

Lietzan, Erika, and Kristina M. L. Acri. 2020. "Distorted Drug Patents." *Washington Law Review* 95 (3): 1317–82.

Mansfield, Edwin. 1986. "Patents and Innovation: An Empirical Study." *Management Science* 32 (2): 173–81.

Mezzanotti, Filippo, and Timothy Simcoe. 2023. "Innovation and Appropriability: Revisiting the Role of Intellectual Property." NBER Working Paper 31428.

NBER. 2023. *Orange Book Patent and Exclusivity Data—1985–2016.* https://www.nber.org/research/data/orange-book-patent-and-exclusivity-data-1985-2016 (accessed September 1, 2023).

Office of the Federal Register, National Archives and Records Administration (NARA). 2023. *Federal Register Documents Search.* https://www.govinfo.gov/app/collection/FR (accessed September 1, 2023).

Olson, Luke M., and Brett W. Wendling. 2018. "Estimating the Causal Effect of Entry on Generic Drug Prices Using Hatch–Waxman Exclusivity." *Review of Industrial Organization* 53 (1): 139–72.

Peltzman, Sam. 1973. "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments." *Journal of Political Economy* 81 (5): 1049–91.

Podolsky, Scott H. 2015. *The Antibiotic Era: Reform, Resistance, and the Pursuit of a Rational Therapeutics.* Johns Hopkins University Press.

Rai, Arti K., Saurabh Vishnubhakat, Jorge Lemus, and Erik Hovenkamp. 2022. "Post-Grant Adjudication of Drug Patents: Agency and/or Court?" *Berkeley Technology Law Journal* 37 (1): 139–70.

Reddy, Sanjay, Reed F. Beall, S. Sean Tu, Aaron S. Kesselheim, and William B. Feldman. 2023. "Patent Challenges and Litigation on Inhalers for Asthma and COPD." *Health Affairs* 42 (3): 398–406.

Roin, Benjamin N. 2009. "Unpatentable Drugs and the Standards of Patentability." *Texas Law Review* 87 (3): 503–70.

Roin, Benjamin N. 2013. "The Case for Tailoring Patent Awards Based on Time-to-Market." *UCLA Law Review* 61 (3): 672–759.

Sampat, Bhaven. 2020. "Whose Drugs Are These?" *Issues in Science and Technology* 36 (4). https://issues.org/drug-pricing-and-taxpayer-funded-research.

Sampat, Bhaven. 2023. *The History and Political Economy of NIH Peer Review.* Brookings Institution.

Shapiro, Carl. 2003. "Antitrust Limits to Patent Settlements." *RAND Journal of Economics* 34 (2): 391–411.

Song, Yan, and Douglas Barthold. 2018. "The Effects of State-Level Pharmacist Regulations on Generic Substitution of Prescription Drugs." *Health Economics* 27 (11): 1717–37.

SSR Health. 2024. *SSR Health Drug Sales Database* (accessed June 25, 2024).

Taylor, C. T., and Z. Aubrey Silberston. 1973. *The Economic Impact of the Patent System: A Study of the British Experience.* Cambridge University Press.

Temin, Peter. 1979. "Technology, Regulation, and Market Structure in the Modern Pharmaceutical Industry." *Bell Journal of Economics* 10 (2): 429–46.

Temin, Peter. 1980. *Taking Your Medicine: Drug Regulation in the United States.* Harvard University Press.

USTPO (US Patent and Trademark Office). 2023a. *Patent Examination Data System (PEDS).* https://ped.uspto.gov/peds/# (accessed September 1, 2023).

USPTO. 2023b. "Patent Terms Extended Under 35 U.S.C. § 156." https://www.uspto.gov/patents/laws/patent-terms-extended (accessed September 1, 2023).

USPTO. 2023c. *PatentView.* https://www.uspto.gov/ip-policy/economic-research/patentsview. Accessed September 1, 2023.

Voet, Martin A. 2005. *The Generic Challenge: Understanding Patents, FDA and Pharmaceutical Life-Cycle Management.* 6th ed. BrownWalker Press.

Wouters, Olivier J., Martin McKee, and Jeroen Luyten. 2020. "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009–2018." *JAMA* 323 (9): 844–53.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 53–78

Lessons for the United States from Pharmaceutical Regulation Abroad

Margaret K. Kyle

American politicians from across the ideological spectrum have found a rare point of agreement: US drug prices are too high, and the government should do something about it. Figure 1 presents the results of a study of drug prices in different countries, which finds that overall US pharmaceutical prices, and particularly prices of on-patent drugs, are several times higher than those in other high-income countries such as France, Germany, Japan, and the United Kingdom (Mulcahy, Schwam, and Lovejoy 2024). Such comparisons bolster the arguments of those advocating price controls or other government interventions in the United States.

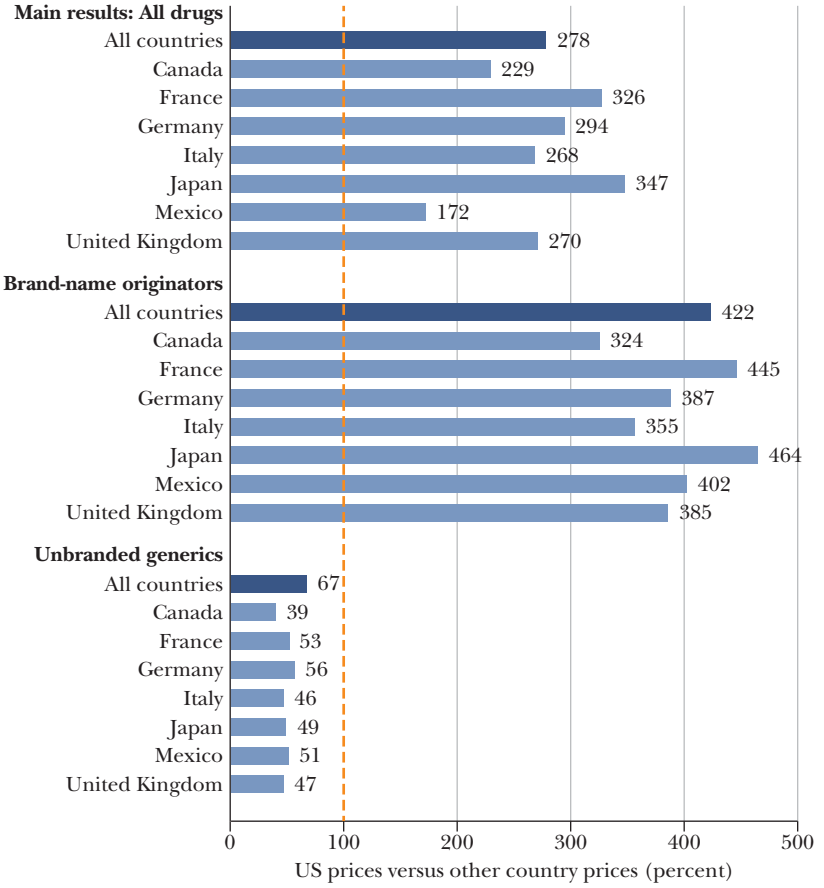
Why are policies around drug prices necessary? In other words, why do most developed countries rarely leave drug pricing to "the market" without government intervention? In all countries, pharmaceutical markets operate under conditions that differ substantially from the assumptions behind a competitive market that would equilibrate supply and demand based on price. These conditions reflect information problems, the specificities of healthcare markets, and the cost structure of drug development.

First, pharmaceuticals are "credence goods": not only is their quality difficult for consumers to determine prior to consumption, but their effects are challenging to identify even after consumption. Regulation of entry, administered by the US Food and Drug Administration (FDA) and similar regulators like the European Medicines Agency (EMA), aims to reduce this information asymmetry. Consumers can be

■ *Margaret K. Kyle is Professor of Economics, Center for Industrial Economics (CERNA), Mines Paris-PSL (Ecole des Mines), Paris, France. Her email address is margaret.kyle@minesparis.psl.eu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241418. The author has received financial support from an interested party.

Figure 1
US Manufacturer Gross Prescription Drug Prices as a Percentage of Prices in Selected Other Countries, All Drugs, 2022



Source: Mulcahy, Schwam, and Lovejoy (2024).
Note: "All countries" refers to all 33 OECD comparison countries combined. Other countries' prices are set to 100. Biologics were excluded from unbranded generics. Only some presentations sold in each country contribute to bilateral comparisons. Brand-name originators and Unbranded generics reflect IQVIA's assignment of drug products in individual countries.

reasonably confident that the little pink pill they swallow is what it says on the label, and that clinical trials have demonstrated the safety and efficacy of the product. Absent such quality regulation, the market may shrink or collapse entirely (while manufacturers could signal quality through other means, such as brand name reputation or the use of certification by third parties, these would also likely raise costs and limit entry). A consequence of regulation, however, is that pharmaceutical markets do not have free entry—even as the structure of the industry has evolved with the entry of many smaller biotechnology and genomics firms.

Demonstrating quality to regulators requires the provision of detailed information from clinical trials as well as manufacturing processes. Development costs are very high for a new drug (DiMasi, Grabowski, and Hansen 2016). However, imitation of a drug whose safety and efficacy have been established is relatively easy, particularly if the detailed information provided to regulators is public. Innovators may not be able appropriate the benefits if rapid entry by competitors occurs. These conditions render investment in new drug development unattractive and explain a second intervention in pharmaceutical markets: patent protection and other forms of exclusivity. By shielding a new drug from direct competition for some period of time, these policies seek to encourage investment in innovation. But this regulatory protection is another reason that free entry does not exist in pharmaceuticals.

If we take entry regulation and patent protection as given, therefore, government intervention yields pharmaceutical markets that are not perfectly competitive. Few markets are, of course, but as a healthcare good, the purchase of pharmaceuticals takes place in a particular context. Specifically, many drugs are covered, at least in part, by health insurance.

Because patients with insurance do not face the true price of care, they may consume too much. Einav and Finkelstein (2018) provide an assessment of empirical studies of moral hazard and health insurance, and conclude that the evidence is consistent with moral hazard. This moral hazard increases costs for insurers, even when treatments are supplied in a perfectly competitive market, by increasing the quantity consumed. When suppliers have market power as a result of patents or other exclusivity, however, the problem is further aggravated. Facing patients with demand made less elastic by insurance, suppliers can set higher prices.

A further complication is that prescribers mediate demand for pharmaceuticals. Like many goods and services in healthcare, experts (usually physicians) are presumed to have better information on the appropriate treatment than the consumer (patient). For prescription drugs, the prescriber selects the drug and dosage. In Western countries, prescribers do not typically sell drugs to patients. They may not be aware of relative prices or be sensitive to them. This is partly by design: a stronger profit motive could introduce problematic agency issues (for evidence on this point, see Jacobson et al. 2006; Iizuka 2007). However, the separation of prescribing and dispensing drugs limits the price elasticity of prescribers, although there is evidence that prescribers do respond to information about price when it is made available (Epstein and Ketcham 2014).

Given all the informational difficulties and regulation in healthcare, prices determined in a market without price regulation may not provide an efficient outcome or profits that correctly incentivize innovation and provision. While regulations reduce asymmetric information, they do not eliminate it. Even if clinical trials provide an estimate of a drug's effectiveness on average, patients and physicians may have difficulty identifying the causal effect of a treatment on the individual. Placebo effects may play a role, for example, or other interventions or patient behaviors may be important. Physicians may not be well-informed about recent drug developments and innovations, or their clinical practices may evolve very slowly; this

can slow the adoption of therapeutically beneficial products (Kyle and Williams 2017). Advertising and marketing efforts may also influence demand, with ambiguous and heterogeneous effects (Leffler 1981; Azoulay 2002; David, Markowitz, and Richards-Shubik 2010; Ching and Ishihara 2012). Due to secret rebates and other factors that reduce transparency, the price of a pharmaceutical product may convey little information about its quality.

In summary, with regulatory barriers to entry, insurance, and prescribers without "skin in the game," conditions are ripe for very high prices on patented pharmaceuticals. This presents a budgetary challenge for payers who act as price-takers. They may respond by restricting access, with negative consequences for patient welfare. If pharmaceutical prices are unrelated to a drug's importance (that is, therapeutic value and effectiveness), the market is not sending appropriate signals for the direction of further investment (Kyle 2018).

Government intervention in pricing can, in theory, improve outcomes. The following sections discuss such interventions in more detail. The United States is an outlier among developed nations in having a more limited role of government in healthcare generally, and in pharmaceutical pricing specifically. While US drug prices are higher, spending on healthcare is also much greater in the United States. In addition, the sheer size of the US market gives its policies a greater role in global innovation incentives.

If a decentralized market has shortcomings, designing a well-functioning centralized market for pharmaceuticals poses numerous challenges. Some of these are common to other healthcare goods and services, like accounting for the opportunity cost of spending on drugs versus other treatments, or health versus other government programs. Administrative costs and the potential for gaming are also issues in setting reimbursement rates for procedures, office visits, and other areas. Numerous studies have examined strategic responses to regulated reimbursement rates, such as "upcoding" or a change in services provided (Dafny 2005; Fang and Gong 2017; Batty and Ippolito 2017). When these fees are vulnerable to manipulation by providers, they may be set at inefficient levels (Milcent 2021).

Other issues are more salient when intervening in pharmaceutical pricing than for most healthcare goods or services. Drug pricing is implicitly a tool of innovation policy. If innovation is not rewarded, we should expect lower levels of investment. Patents are ineffective policy instruments for pulling investment in innovation if prices are determined by maximizing only *static* welfare. The cost structure of pharmaceuticals is important here, because prices are usually determined after the costs of development have been incurred. If firms sell to monopsonistic purchasers, they are subject to hold-up. Purchasers then must commit to prices that exceed what is optimal in the short run. In addition to providing incentives for the level of innovation, prices must also provide information about the desired direction of innovation. That is, they should align social value and private rewards so that investors pursue innovation that is socially valuable.

Of course, measuring the social value of a pharmaceutical product is far from simple. The clinical trials required for market authorization by the US Food and

Drug Administration or its counterparts in other countries provide information about a drug's safety and efficacy under carefully controlled conditions; they do not, however, assess overall therapeutic value or social value more broadly. For example, a clinical trial may show that a cardiovascular treatment reduces heart attacks by 10 percent. But that does not tell us whether the treatment is more effective in an absolute sense than existing treatments, or whether it is more cost-effective, or whether that reduction leads to lower spending on emergency care or surgical interventions, or more productive days in employment. The market design for pharmaceutical price regulation must create incentives for the provision of this information, or incur the administrative costs of producing it.

Finally, unlike most services in healthcare, pharmaceuticals are globally traded products that are very similar across countries. The possibility of price arbitrage between markets may result in firms choosing to forgo smaller markets with low prices in order to preserve profits in large countries (Kyle 2007; Maini and Pammolli 2023). In addition, the innovation incentives created by any single country spill over to all other countries that might benefit from the results of that innovative effort. Small markets that account for a small share of global revenues may see little reason to create innovation incentives through paying high prices, as the private sector is unlikely to respond strongly to a small increase in expected revenues. Large markets like the United States, which do have the potential to shift innovation incentives through prices, are thus subject to free-riding (Grossman and Lai 2008). The inclusion of a minimum level of intellectual property protection in trade agreements seeks to limit that free-riding, but does not eliminate it if countries use price controls. These international externalities introduce an important complexity to market design.

The Landscape for Setting Drug Prices

Developed countries around the world have introduced an array of policies to control pharmaceutical prices. I focus here on those used for on-patent or branded drugs, although these have implications for the generic market as well. Some countries intervene directly in markets by setting a maximum price of a product, at any or all of the manufacturer, wholesale, or retail levels. Others allow manufacturers to set prices in response to government policies on coverage and reimbursement. Countries vary in how they evaluate products; for example, on an absolute or comparative basis, and in whether they consider only clinical characteristics or other factors, such as research and development costs or local production. I do not attempt a comprehensive survey here; useful summaries are provided by the Commonwealth Fund (2020) and the OECD (2008).

Types and Basis of Intervention

As monopsonist buyers in most countries, government payers have considerable sway over the price of drugs, but they have adopted different approaches. In

Table 1
Overview of Reimbursement and Pricing Regulations, Selected Countries

	Reimbursement decision			*Factors in price-setting*			
Country	*Timing*	*Economic considerations*	*Price regulation*	*Cost-plus*	*Therapeutic referencing*	*International benchmark*	*Profit regulation*
Australia	Before launch	Yes	Maximum ex-factor price	Yes	Yes		
France	Before launch	No	Maximum statutory price		Yes	Yes	
Germany	After launch	No	Statutory price negotiated after market entry		Yes	Yes	
Japan	Before launch	No	Reimbursement price		Yes		
United Kingdom	After launch	Yes	Not direct				Yes

Source: Paris and Belloni (2013).

some countries, prices are the result of a direct negotiation between the government and the manufacturer. Other countries take the manufacturer's price and decide whether the government will reimburse it. With either approach, governments use a variety of criteria to value a new drug; Table 1 provides an overview for several OECD countries.

Evaluation Criteria

Cost-based approaches. "Cost-plus" or "rate-of-return" regulations, like those once widespread in electricity markets, appeal to those who believe pharmaceutical firms earn excessive profits. The pharmaceutical industry itself uses the high costs of research and development as a justification for its pricing. Such policies can be applied at the product level, setting the price of a drug to its estimated development costs plus a mark-up, or at the firm or product portfolio level, in which the firm's overall rate of return is limited.

The UK Pharmaceutical Price Regulation Scheme (PPRS), now known as the Voluntary Scheme for Branded Medicines Pricing, is an example of rate-of-return regulation. In use since 1957, it is an agreement between the UK government and the "research-based pharmaceutical industry," members of the trade group Association of the British Pharmaceutical Industry (ABPI), and is renegotiated every five years. Participating manufacturers are subject to a maximum profit (currently 21 percent) allowed on their sales to the National Health Service (NHS); they also agree to limit growth in the total spending by the NHS on products covered by this scheme. Profits in excess of this cap are paid back to the NHS as rebates. Note that there is no minimum profit guarantee, and that the profit controls are assessed on a portfolio of products rather than drug-by-drug. Little empirical evidence exists on the effects of this policy. Borrell (1999) assesses outcomes through 1994, and concludes that the PPRS had little effect on prices. The ABPI (2014) documented a decline in the relative prices of branded drugs in the United Kingdom compared

to other countries between 2003 and 2011. Implications for the level and direction of research and development investment are even harder to identify because the UK market is a small percentage of global revenues.

Economists are generally skeptical of cost-plus or rate-of-return regulation in this context, particularly if applied to individual drugs rather than product portfolios. Cost-plus, in particular, risks rewarding research and development spending that may be inefficient. Firms are willing to spend more to develop products with higher expected revenues, and additional spending can result in a better clinical profile. However, more money spent on developing a product does not always yield greater therapeutic value, for the same reason that millions spent on producing a film does not guarantee either its commercial success or its artistic value. Setting a higher price for a drug that was expensive to develop but that has the same clinical effects as one with low development expenses both increases costs for the payer as well as distorting incentives. Rate-of-return regulation applied to a portfolio of products creates no incentive for firms to bring development costs down; these have increased substantially over time (DiMasi, Hansen, and Grabowski 2003; DiMasi, Grabowski, and Hansen 2016), potentially representing an important barrier to entry as well. With both cost-plus as well as rate-of-return regulation, price is not necessarily linked to therapeutic or social value. In the former case, as just discussed, development cost and therapeutic value may not be related. In the latter, if firms may freely price, the link with social value may be weakened due to market frictions discussed in the previous section.

A second concern is the need for detailed internal data from firms on their expenses in order to determine costs. The World Health Organization recommends against cost-plus regulation in its 2020 Guidelines mainly for implementation challenges, noting that the "[a]pplication of cost-plus pricing to medicines requires significant technical and human resources, particularly in obtaining and validating reliable estimates of component costs." The difficulty of estimating costs may be one reason that cost-plus policies are not widespread in developed countries. Recent calls for increased transparency in research and development spending, such as spending by product rather than the aggregate figures usually reported, target this difficulty. Rate-of-return policies at the firm level have a lower, but still important, information burden.

Value-based criteria. In recent years, the industry has advocated the adoption of "value-based" pricing, in which research and development and manufacturing costs play no role. Rather, in this approach, more valuable (therapeutically superior) products receive higher prices. Because a product may be more effective for certain patient populations, or may treat several diseases with varying effectiveness, a value-based price may differ for the same drug across therapeutic indication or patient subgroup, at least in theory.

There are at least two challenges to the implementation of value-based pricing. First, while it does not require information on research and development costs, it does require assessments of a drug's therapeutic value that go beyond what is

typically required for market authorization. These health technology assessments often include comparative studies, in which a trial tests a new product against other treatments available rather than a placebo. The clinical endpoints may differ, and these studies may exploit "real world evidence" to examine the use of a product outside the carefully controlled clinical trial setting. These assessments are often costly and time-consuming. While regulatory agencies have largely harmonized requirements for marketing authorization, there is less coordination across countries in designing health technology assessments, and this can lead to different conclusions about a drug's value. In part, that stems from the variation in standard of care or the importance of competing products, which are not the same in every country. Satisfying different demands for health technology assessments adds to the manufacturer's costs, although manufacturers may prefer a diversity of views on a drug's value if it allows them to price discriminate across countries. A second challenge is that if therapeutic superiority can be established, it still remains to determine how much a payer is willing to pay for that clinical improvement.

In France, the pricing and reimbursement process begins with an initial determination of whether a product is important enough to be reimbursed (its "service médical rendu" (SMR)) and its therapeutic advantages over existing treatments (ASMR or "amélioration de service médical rendu"). These assessments are officially independent of price; the ASMR is an input into subsequent price negotiations, which typically establish a price for a three-year period. Products judged to add therapeutic value are entitled to a price premium over competitors, while others must be priced at or below existing products. However, perhaps in anticipation of those negotiations, the body assigning ASMRs has historically judged very few products to add significant therapeutic value, and the revenues of products are only loosely related to their ASMR scores (Kyle 2018). Although a drug may have several ASMRs, varying by indication or patient subgroups, there is a single price established for the entire market, illustrating the practical difficulty of indication-based pricing linked to value.

Cost-effectiveness thresholds. An approach that avoids haggling directly over price instead has payers exercise their power to refuse insurance coverage. For the reasons discussed above, in most health systems, insurance reduces the price elasticity of patients, and prescribers are unaware of or insensitive to the price of drugs. If manufacturers are free to set price, full reimbursement of all drugs is inefficient and impossible from a budgetary standpoint. Payers can therefore choose to limit coverage by reimbursing only those products whose prices are less than the payer's willingness-to-pay for a health improvement.

The United Kingdom and Germany both use cost-effectiveness evaluations as the main policy instrument to address price. Manufacturers can freely set prices of their products in the United Kingdom; these products are subject to cost-effectiveness evaluation by the National Institute for Clinical Excellence (NICE), which decides whether to recommend a product for reimbursement by the National Health Service. Germany allows the manufacturer to establish a price at the time the product is launched, and six months later, evaluates whether the

country will reimburse the product at that price or at the price of a competing product. Of course, the manufacturer determines its price in the shadow of the reimbursement decision.

Given a manufacturer's chosen price, a health technology assessment evaluates the product's cost per quality-adjusted life year (QALY) and determines whether the product should be reimbursed. A QALY is based on a combination of life expectancy and quality of life, where quality of life is measured on a scale from zero to one, from death to perfect health. The calculated QALY for a given drug can vary across countries depending on the data and methods used. This willingness-to-pay can be explicitly announced, or governments can provide guidance on a range of thresholds. Prices are not negotiated directly with payers, but manufacturers anticipate the need to adjust price in response to cost-effectiveness thresholds.

The health technology assessment (like that done for value-based pricing) is a large information burden. Cost-effectiveness policies also require consensus on the value of a statistical life. Allowing for different valuations by disease or target population is theoretically possible, but may be difficult to justify politically. Thresholds used in practice vary by country; the King's Fund in the United Kingdom estimates that the incremental cost-effectiveness ratio threshold in the United Kingdom is lower than that used in Ireland and the Netherlands, for example (Collins 2020).

The case of Ocrevus (ocrelizumab) illustrates the variation in health technology assessments across countries, with implications for price, reimbursement, and access. The French assessment concluded that it offered moderate therapeutic benefits for primary progressive multiple sclerosis (PPMS) and minor benefits for relapsing-remitting multiple sclerosis (RRMS), and the drug is reimbursed for both. The German assessment also determined minor benefits for RRMS, but found the evidence insufficient for PPMS. In the United Kingdom, the National Institute for Clinical Excellence has recommended it for PPMS, but only for RRMS if another treatment is contraindicated or unsuitable. Each country may put different weight on the strength of the evidence, value some patient subsets more than others, choose different comparator products, or focus on different clinical outcomes.

These policy choices, with their varied informational and administrative burdens, have implications for the speed of access to new drugs. Manufacturers launch new products in Germany almost immediately after obtaining market authorization. The administrative steps in countries like France, Italy, and Spain contribute to months-long delays of product launch. It is certainly possible that the negotiation process could be streamlined or accelerated. However, launch delays can also be strategic, as discussed below.

Other factors. Other criteria are also factors in negotiations over price or the level of reimbursement. Paris and Belloni (2013) provide an excellent description of how different OECD countries approach this. For example, some governments may reward manufacturers for investing in research and development or manufacturing locally. Orphan drugs, which treat small patient populations, may be subject to different reimbursement rules. Often, both price and quantities are negotiated

simultaneously. For example, firms may commit to price reductions if volumes exceed a certain level. In assessments that consider more than clinical performance, payers must consider what offsetting costs or benefits to account for. A therapy that reduces absenteeism or increases productivity generates social benefits, but may not reduce costs directly incurred by the health system. Valuing such benefits may then favor treatments for younger patients, rather than the elderly. Increasingly, countries are experimenting with alternative approaches, such as those described later in this paper.

Pricing Benchmarks and Parallel Trade

In well-functioning markets, prices convey information. Government interventions in drug pricing often use other prices as benchmarks, as a complement to the procedures discussed above. These comparisons can be to other drugs that treat the same condition, or to the price of the same drug in other markets. These two types of "reference pricing" have very different effects, however. Governments can also allow for prices in other countries to affect local prices through policies on parallel trade or reimportation.

Internal reference pricing. Internal reference pricing establishes a new product's reimbursement level as equal to that of a competing product, or reference product, within the same country. In some countries, manufacturers are free to set a higher price, but patients must then pay the premium over the reference product price. This has some similarities to tiered formulary pricing in the US context. It introduces some price sensitivity and allows the market to reward quality. For example, if patients value an improvement in dosing (for example, once a week instead of daily, or an oral formulation instead of injection) or have a strong preference for branded products over generics, then manufacturers should arrive at a price that reflects that valuation. However, the efficiency of this policy relies on how well-informed patients and prescribers are about the clinical characteristics of competing products and on the appropriate choice of a reference product.

Experiences with the use of internal reference pricing vary across countries. While Brekke, Holmas, and Straume (2011) and Kaiser et al. (2014) find that its introduction in Norway and Denmark increased the use of generic substitutes and lowered average prices, the policy led to higher patient co-payments and an increase in the price of off-patent products in Portugal (Costa and Santos 2022). Depending on how the reference price is calculated, the effects are theoretically ambiguous (Miraldo 2009). Because these policies have the potential to affect revenues—and in particular, the distribution of revenues between innovative and generic products, or through rewards to quality—they could also affect innovation incentives. However, isolating the effect of a single country's adoption of such a policy on global investment is a challenge.

External reference pricing. In contrast, external reference pricing compares the price of a drug to the same product in other countries. Many European countries

use some form of this policy. For example, in 2012, the Netherlands limited the price of a new drug to be no more than the average in Belgium, France, Germany, and the United Kingdom.[1]

An economic justification for this benchmarking could be that if the referenced countries have useful health technology assessments of a new product and that information is reflected in their prices, there is no need for another country to perform one—it can instead rely on the information generated by other payers. This is an argument for coordination on a single health technology assessment. Because external reference pricing is often used alongside a country's own health technology assessment, another economic argument is that it is like a meta-analysis of comparative studies, filtered through pricing. However, the price in the reference country is likely to be a function of more than just the health technology assessment. Finally, small countries may reference prices in larger markets to counteract their lower bargaining power with manufacturers—but small markets like Belgium, Denmark, and Ireland are often referenced by larger countries. As implemented, therefore, the main motivation for external reference pricing seems to be the perception of fairness: why should a country pay higher prices than its neighbors?

An important argument against external reference pricing is that it creates incentives for strategic launching of new products to the detriment of countries with relatively low prices, if those countries are used as external references. If markets were completely independent and manufacturers faced no supply constraints, pharmaceutical firms should be willing to sell at any price that covers their marginal costs and country-specific fixed costs. However, most drugs are not launched immediately in every country, or even all developed markets. In addition to the bureaucratic delays associated with pricing and reimbursement, there is an incentive for manufacturers to delay launch (or never launch) in referenced countries that negotiate low prices. Because the price in a referenced country affects the price in the referencing countries, manufacturers favor faster launch in countries with higher prices. In practice, those are often the richest countries; these strategic launch delays are costly to those with lower income levels and whose prices are referenced (Kyle 2007; Maini and Pammolli 2023).

Parallel trade in pharmaceuticals (or reimportation, as it is sometimes called in the United States) creates similar incentives. If permitted, drugs sold in other countries at a lower price can be imported to compete with local products. This arbitrage of cross-country price differences makes launch in low price markets less attractive, along with other consequences like shortages (Kyle 2011). Both external reference pricing and parallel trade limit the ability of firms to price discriminate across countries. However, economists often see price discrimination as a relatively efficient outcome, particularly if access to new products in poorer countries is important.

Neither external reference pricing nor parallel trade has entirely eliminated cross-national price differences. Importantly, prices are not simply a function of

[1] For details, see Figure E.3 of Maini and Pammolli (2023).

GDP per capita, but reflect the policy choices, preferences, and bargaining power of countries. For example, injectable semaglutide (Ozempic) was priced at $83 in France, but $103 in Germany; in tablet form (Rybelsus), it was $203 in the Netherlands, compared to $103 in Sweden (Amin et al. 2023). Launch delays and access are also stark. Of the drugs approved by the European Medicines Agency from 2019 to 2022, 88 percent were available in Germany by early 2024, but less than 20 percent for the Baltics and Romania (EFPIA 2024).

Can the United States Import Similar Policies?

While governments are the main health payers in other developed countries, the private sector has a larger role in the United States, and drug coverage is not universal. In other developed countries, most patients face approximately the same price for a drug. In the decentralized US system, however, the same drug may be sold at very different prices to different buyers. Pharmacy benefit managers or the insurance companies who use them negotiate pharmaceutical prices with manufacturers. While no pharmacy benefit manager is a monopsonist, they do control access to large patient populations, and they are not price-takers (Guardado 2024). Manufacturers negotiate both price and reimbursement (or formulary placement) with pharmacy benefit managers. As with government payers, it is important that pharmacy benefit managers act as good agents for patients. One critical difference with national payers is that even if US patients face switching costs in changing insurers, residents of a country are less likely to "vote with their feet" by moving abroad if dissatisfied with pharmaceutical pricing and access. That is, pharmacy benefit managers face some competition and may lose market share if they refuse to cover an important drug. A second important difference is that pharmacy benefit managers try to maximize profits, not social welfare. The influence of any single pharmacy benefit manager on the level and direction of innovation is limited, but they also may not realize the long-run benefits of innovation to the same extent as a government payer. For example, the benefits of a childhood vaccine accrue to a government payer over an individual's entire life, in the form of reduced illness and associated healthcare costs. But that individual is unlikely to be the client of the pharmacy benefit manager over their entire lives, nor is the pharmacy benefit manager necessarily sensitive to the reduction in associated healthcare costs.

The US federal government is also a major payer, but its interventions in drug pricing until recently were limited. While the Department of Veterans Affairs can negotiate directly, it is a very small share of government spending on drugs. The price Medicaid pays is linked to that in the private sector, but not negotiated directly. The 2003 Medicare Modernization Act barred the federal government from bargaining over prices for drugs reimbursed under Medicare Part D. Instead, the private sector (insurers who offer Medicare Part D and pharmacy benefit manager) takes that role.

This has recently changed. The Inflation Reduction Act of 2022 introduced the Medicare Drug Price Negotiation Program, which establishes a Maximum Fair Price (MFP) on single source products with large shares of Medicare spending whose active ingredients were first approved at least seven years ago (or eleven in the case of biologics). Prominent health economists argue that this negotiation should be based on value, linking prices to measures of clinical and economic benefits (Conti, Frank, and Gruber 2021). As adopted, however, the law mandates price discounts based on the number of years since US launch. Many unintended consequences may result due to strategic responses by manufacturers. For example, although a limited patent term generally encourages firms to launch as quickly as possible, some have suggested that that firms may avoid launching early in a single indication to avoid triggering the "clock" for Medicare negotiations (Patterson, Motyka, and O'Brien 2024). Mandatory price discounts in advance of generic or biosimilar entry have the effect of reducing the incentives for that entry, which usually responds to the size of a drug's revenues (Danzon and Chao 2000). Firms may also have more incentive to "product hop," which refers to the introduction of a new version of an existing drug prior to patent expiration. Long a concern to competition authorities (Federal Trade Commission 2022), this can lead to substitution away from the product subject to Medicare negotiation and towards a more expensive product, reducing the expected cost savings from the policy. A crucial point is that in contrast to most of the countries discussed above, the United States is a big enough market for its policy choices to "move the needle" on research and development investment. Total spending on pharmaceuticals in the United States is almost four times higher than in China, five times higher than in Japan, and ten times higher than in Germany (IQVIA Institute 2024). The balancing of static and dynamic effects is therefore more important than in the case of a country the size of Belgium or Ireland. An explicit policy goal to reduce total spending is likely to reduce investment, and imperfect assessments of value (for example, through erroneous assumptions used in a health technology assessment or discounts that are independent of value) may distort it. The magnitude of these effects is the subject of intense debate.

The incentives created by the United States historically have allowed smaller markets to free ride, at least to some extent. This has prompted proposals to link US prices to those in other markets. Reimportation of drugs from Canada or other developed markets has been proposed, but not yet implemented, at the federal level. In 2019, the Trump administration proposed creating pathways for reimportation of pharmaceuticals (FDA 2019). Some states have moved in this direction, including Florida and several states in the Northeast. Their legal authority to do so has been challenged, but the US Food and Drug Administration approved a plan to allow reimportation in early 2024. The consequences of parallel trade in Europe suggest that reduced access to new drugs in Canada is a serious risk. Canada would also likely take steps to limit export in order to protect access for Canadian patients. The relative size of the Canadian market (approximately one-tenth that of the United States) implies that even if every product were reimported into the US market, the change in the overall US price level would be small (Kyle 2011).

Tying US prices to prices in other countries—that is, external reference pricing—is also a policy proposal with bipartisan support. Though not implemented, the Medicare Most Favored Nation pilot program would have capped prices at the lowest paid by other countries with at least 60 percent of US GDP per capita (Centers for Medicare & Medicaid Services 2021). An optimistic take on this applies the ideas of Grossman and Lai (2008), whose model predicts that prices in other countries would increase (as would research and development and firm profits) as a result of this linkage. Industry trade groups strongly oppose this policy, though, suggesting that manufacturers expect a negative effect on firm profits.

A less optimistic assessment of external reference pricing considers the European experience. As noted above, external reference pricing like this would induce a number of strategic responses from other stakeholders. These include delayed launch and/or supply limitations to lower-price markets, as well as efforts to make products less comparable across countries (Kyle 2007, 2011; Maini and Pammolli 2023). Another response is similar to the approach taken by pharmacy benefit managers, who negotiate secret rebates with manufacturers. That is, the true price paid by pharmacy benefit managers is not easily observed; there is a list price, but the rebate or discount offered by the manufacturer is not public. Some European countries also use hidden rebates. For example, the use of France as a reference by other countries ultimately led to agreements between manufacturers and the government to establish a public price as well as secret rebates paid by manufacturers back to the government (Kanavos et al. 2017). This allows the official price (that which is referenced by other countries) to be higher, like the list price in the United States, than what is in fact paid. These nonpublic prices have prompted calls for greater price transparency, but the effects of increased transparency here are ambiguous. When (true) prices are secret, a manufacturer can more easily lower its price in a country, because it sees no negative consequences from having that secret price referenced by other countries. In concentrated markets, transparent prices could also facilitate collusion by manufacturers. However, nonpublic prices make economic assessments much more challenging.

The evidence suggests that US adoption of reimportation or external reference pricing would have only modest effects on US drug prices (but would probably reduce access or price transparency in other countries). Negotiating prices paid by Medicare has larger potential effects on the level of US prices, depending on what criteria are used. Importantly, unlike pricing interventions in most countries, this policy could also have large effects on global research and development due to the size of the US market. Numerous studies have shown that research and development responds to increases in market size that result from US government purchases or mandates without price controls, though the innovation is not necessarily socially valuable (Finkelstein 2004; Dranove, Garthwaite, and Hermosilla 2022). The new Medicare policy could, in theory, better align profits with social welfare than the current system and improve incentives. However, as discussed previously, reliable health technology assessments are essential for this purpose. The effects also depend on how private payers respond.

Alternative Approaches

Many stakeholders, both in the United States and other countries, are unhappy with the drug pricing status quo. Payers face budgetary challenges that are most acute when arguably the most valuable therapies reach the market. Sovaldi (sobosfivir) in 2013 is an example: launched at an initial price tag in excess of $80,000 in the United States, it was a large advance in treatment (indeed curing) hepatitis C (HCV). While it was cost-effective even at its list price, the size of the potential patient population caused an abrupt increase in expenditures by payers on treating HCV. Although both private payers in the United States and public payers elsewhere negotiated substantially lower prices, the impact on pharmaceutical spending was nevertheless large. Iyengar et al. (2016) calculated that providing Sovaldi or similar antivirals to all patients with HCV would amount to an increase of more than 10 percent of total pharmaceutical spending in 30 countries studied; for some countries, treating all HCV patients would require doubling total spending on drugs. In the United States, public programs also saw an large increase in outlays. Medicare spent $8.2 billion (before rebates) on HCV treatments in the 18 months after Sovaldi's launch, six times more than spending on HCV previously. The recent introduction of effective treatments for obesity for a larger potential population could strain public budgets even more. According to Deese, Gruber, and Cummings (2024), "[u]nder reasonable assumptions and at current prices, making this class of drugs available to all obese Americans could eventually cost over $1 trillion per year."

Some countries are considering or experimenting with alternatives to the regulatory approaches outlined above. Many of these rely primarily on patents as the policy tool for creating innovation incentives. More significant departures from the status quo would involve ex ante commitments (that is, before a treatment has been developed).

Contracting before Innovation

In general, countries face a trade-off between providing incentives for innovation through high prices and providing access to that innovation. As discussed above, the many frictions in establishing drug prices and ensuring appropriate use—whether in the United States or other developed countries—yield profits that are not aligned with social value, distorting incentives. De-linking rewards for innovation from high-margin sales, and instead using fixed payments for successful innovation, is an attempt to avoid this trade-off.

Kremer and Williams (2010) suggest experimentation with alternatives to intellectual property rights, including prizes or advance market commitments (AMCs). These mechanisms have payers announcing a reward for developing a treatment meeting specified criteria, and have many appealing features. Because price is not negotiated after the product reaches the market, the manufacturer faces a reduced risk of hold-up, and funders have more certainty about budgets. There is no need to ration treatments to patients, as was the case for hepatitis C drugs, or for

manufacturers to encourage inappropriate use. While the patent system is a blunt policy tool, as the length of protection is not a function of therapeutic value or importance and prices may be hard to predict, a prize system provides much clearer innovation incentives.

One prominent example of an advance market commitment is the Gates Foundation's $1.5 billion pilot financing for pneumococcus vaccines in 2009. More recently, Operation Warp Speed used AMCs for the development of COVID-19 vaccines (Slaoui and Hepburn 2020). Evaluations of these provide valuable lessons for their design and deployment going forward. In particular, both required a number of assumptions, with considerable uncertainty at the time funds were committed, and some of which were criticized afterwards (GAVI 2021; D'Souza et al. 2024). Indeed, this informational burden probably best explains why AMCs are rare. Concerns over the challenges of implementing AMCs or prizes have also been a factor in the failure to adopt the Medical Innovation Prize Fund Act, proposed by Senator Bernie Sanders (I-Vermont) many times.

The use of prizes or advance market commitments also suffers from the same risk of international free-riding as the status quo. In theory, contributions to financing a prize or AMC could be linked to GDP per capita. However, could funders credibly commit to denying access to a resulting treatment to countries that did not contribute?

Contracting after Innovation

Policy changes that leave the patent system largely in place are likely to be easier to implement, because trade agreements such as the Trade-Related Aspects of Intellectual Property Rights Agreement require that signatories provide a minimum level of protection for intellectual property. Patents align private and social value when profits are linked to therapeutic quality and need. Price regulations (or limiting quantities sold through access restrictions) that break this link distort innovation incentives as well as consumption in the short run.

As an alternative to negotiations or "take-it-or-leave-it" offers, auctions are commonly used in government procurement. Indeed, many payers use procurement auctions like this for markets with intramolecular competition. In the United States, for example, the Department of Veteran's Affairs and Group Purchasing Organizations use competitive bidding to procure supply of generic drugs. For on-patent products, this approach is feasible when there are several molecularly distinct treatments that are nevertheless reasonably close substitutes. This is the case for many vaccines, but rarely observed for other situations of intermolecular competition.

In theory, the manufacturer of an on-patent product could itself auction off access to its treatment, or to the associated intellectual property. Kremer (1998) proposes the uses of patent buyouts, using auctions to elicit information about the value of the treatment. In Kremer's proposal, the government would put the intellectual property in the public domain. Manufacturers may be hesitant to participate in patent buyouts if there is uncertainty about the ability to enforce intellectual

property rights in other countries where there was no buyout; organizing a global patent buyout presents the same challenges as international funding of prizes. While the Biden administration recently pressured the National Institutes of Health to exercise "march-in" rights (Department of Health and Human Services 2023), even these are rarely exercised, and I am aware of no government buyout of a pharmaceutical patent.

At least two incremental adjustments to existing pricing approaches could be welcomed both by industry and by payers. Sood et al. (2018) provide an overview of the so-called Netflix model, subscription pricing, or all-you-can eat pricing. In this scenario, there is a fixed payment to the manufacturer for supplying any quantity demanded at zero price (or close to). It is similar, therefore, to buying the patent rights for some period of time; the difference with a patent buyout is that the buyer does not put the intellectual property in the public domain. This approach provides greater certainty for planning budgets, both for payers and for manufacturers. The payment could be staggered over several years, in order to smooth outlays. Because the manufacturer's profit does not depend on the number of units sold, there is reduced incentive to market the product or encourage its use in patients who are unlikely to benefit from it. The payer also should seek to maximize patient access.

The imperfect information at the time of a drug's launch is the same for subscription pricing as for the more traditional approaches described above. While a multi-year agreement provides certainty to both parties, it also makes adjustments to either the arrival of new information or the arrival of new competitors more difficult. This contract also limits the incentive for a manufacturer to disseminate information—while marketing is often criticized, it can play an important role in educating both physicians and patients about the existence of treatments. Manufacturers may also have less interest in finding new therapeutic uses, or improved formulations, if they expect no increase in profits under the subscription model.

Australia implemented this approach in 2015, paying $766 million to four manufacturers of hepatitis C treatments for unlimited access to these drugs over five years. Analysis of the Australian experience suggests that access increased, with a per-patient cost well below the estimated price that would result from negotiations. The Medicaid programs of some US states, including Louisiana and Washington, also negotiated subscription pricing for HCV. Their experiences illustrate that while access has improved, pricing is not the only barrier (Conti, Frank, and Gruber 2021). For example, screening patients who need treatment is itself a challenge. Neither state achieved their initial goals for the number of patients treated.

Pay-for-performance, outcomes-based, or risk-sharing contracts[2] are another potential improvement over traditional pricing. Instead of linking price to therapeutic value demonstrated in trials or studies available at the time a drug is brought to market, these contracts link payments to the realization of *future* clinical milestones or other metrics. The importance of information revealed after market

[2] Many other terms are used for these arrangements; see Towse, Garrison, and Puig-Peiró (2011) for more details.

approval is clear. Prasad et al. (2013) conclude that "[t]he reversal of established medical practice is common and occurs across all classes of medical practice." Many follow-up studies of oncology treatments, which are increasingly approved based on secondary or surrogate endpoints, do not confirm the benefits suggested by earlier clinical trial results (Prasad et al. 2013; Haslam et al. 2021; Mooghali et al. 2024). Boyle et al. (2021) find that "real world data" currently provided on the use of cancer drugs are too sparse or of too little quality to be of use in reimbursement decisions.

The appeal of pay-for-performance contracts is obvious to economists. If well-designed, they address the imperfect information available at launch, but do not delay access while additional information is produced; they buttress incentives for manufacturers to develop treatments with demonstrable gains to patients; and they reduce incentives for manufacturers to encourage the use of a drug in patients who are unlikely to benefit. Better performing drugs generate higher profits, which sends the correct signals for investment in innovation.

Of course, the details in risk-sharing contracts matter. Barros (2011) argues that over-treatment of patients is a risk, and that welfare effects are ambiguous, for the specific contract he models. Contracts that are theoretically well-specified may be impossible to write in practice. Indeed, there is substantial heterogeneity in contract design, to the extent such information is available. (In a survey of the literature on outcomes-based contracts from 2000 to 2019, Antonanzas et al. (2019) note a lack of transparency.) The performance measure can be at the level of an individual patient or more aggregated. While they have the advantage of allowing earlier access despite a lack of complete information about therapeutic value, both parties must agree on appropriate outcome measures. It must be possible to monitor performance, as well as confounding factors that might affect that measure. For example, if a patient adjusts behavior in response to the belief that a drug is effective, that behavioral response could interfere with the measured performance.

In 2002, the United Kingdom's National Health Service established a risk-sharing scheme for four multiple sclerosis treatments that were deemed not cost-effective. Under this program, 5000 patients were monitored over a ten-year period. The agreement was that the manufacturer of a drug that failed to achieve the benefits suggested by initial clinical trials would reimburse the NHS. Initial results were disappointing, and illustrated the difficulties in monitoring patient outcomes (Boggild et al. 2009). However, later results confirmed that the treatments were cost-effective, and are recommended by NICE (National Institute for Health and Care Excellence 2018). Other examples of risk-sharing include the cancer treatment Velcade (bortezomib) in the United Kingdom and Rasilez (aliskiren) for hypertension in Italy (Garrison et al. 2013).

The challenges of negotiating these contracts have hindered their adoption. Payers may also be reluctant because the evidence on the results of pay-for-performance contracts is scant (Antonanzas et al. 2019). Neumann et al. (2011) also point to a lack of data infrastructure and difficulties

in measurement. The absence of a national patient registry and fragmented payers may also hamper their use in the United States. Advances in technologies that enable real-time monitoring of patient compliance and outcomes should reduce the costs of implementing pay-for-performance, if privacy concerns can be overcome.

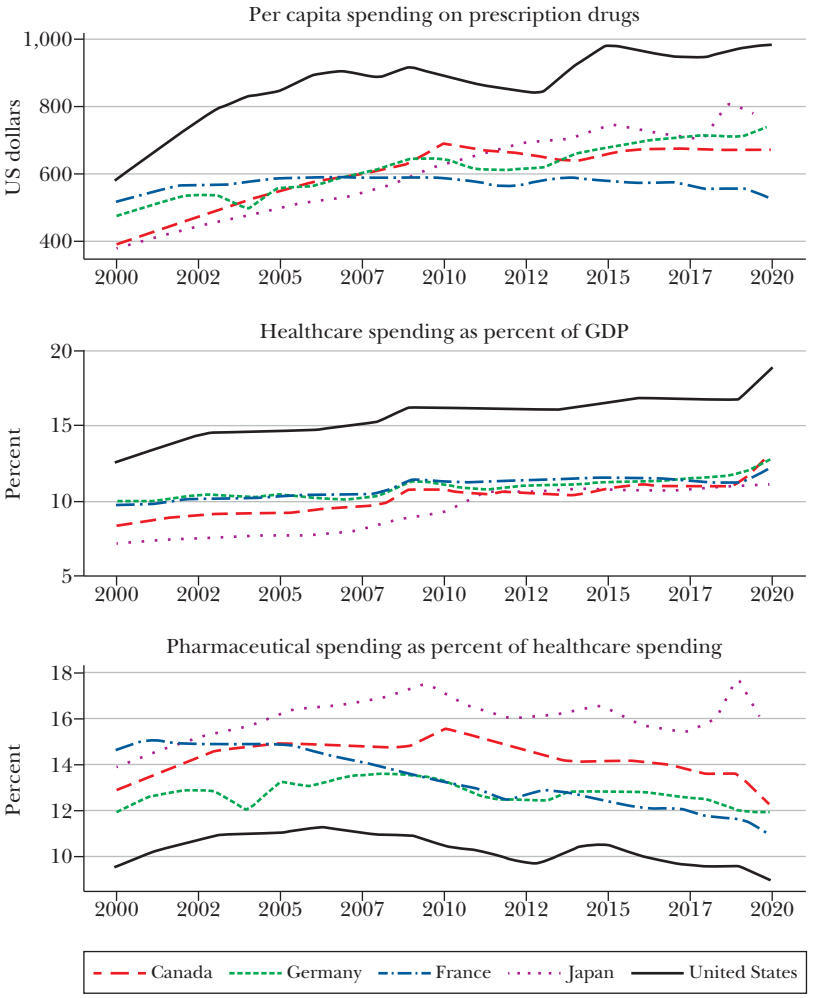
Recent surveys by consulting firms in the United States suggest interest in and adoption of outcomes-based contracts (Chatterjee et al. 2017; Avalere 2023). An increase in the use of risk-sharing contracts, both in the United States and elsewhere, is also noted in Carlson, Chen, and Garrison (2017), Yu et al. (2017), and Piatkiewicz, Traulsen, and Holm-Larsen (2018). However, there is little public documentation of the outcomes of these arrangements. In particular, we lack systematic evidence on whether the cost-savings or health improvements realized offset the difficulties of negotiating them.

The Broader Context of Pharmaceutical Spending

Despite the attention drug prices receive, pharmaceuticals do not seem to be a major driver of healthcare cost increases overall, in the United States or other high-income countries. Drugs are but one input into the broader healthcare production function. The top panel of Figure 2, based on OECD data, shows that per capita spending on prescription drugs is indeed much higher in the United States than in other high-income countries—though this gap is smaller than that for drug prices, as shown in Figure 1. However, US spending on health overall is also much higher: as the second panel of Figure 2 illustrates, health care spending is almost 18 percent of GDP in the United States, but closer to 11–12 percent in Canada, France, Germany, and Japan. The share of pharmaceuticals within health spending is, in fact, lowest in the United States. While healthcare spending as a percent of GDP was essentially unchanged during the pre-pandemic period from 2015 to 2019, the share of pharmaceuticals in that spending declined.

One interpretation is that the prices of other healthcare goods and services in the United States are relatively more expensive than pharmaceuticals. For example, a recent survey of physician salaries across countries in 2021 reported an average of $316,000 in the United States versus $183,000, $138,000 and $98,000 in Germany, the United Kingdom, and France, respectively (Medscape 2021). Similarly, according to OECD Health Statistics, the average salary for hospital nurses in the United States in 2018 was $77,760, versus $56,157 in Canada, $55,302 in Germany, $42,925 in the United Kingdom, $42,099 in Japan, and $41,713 in France. The United States has higher GDP per capita than other countries, and cost-effectiveness thresholds based on the value of a statistical life will naturally allow for a higher US price than in Germany, for example. If a pharmaceutical treatment substitutes for another type of care (for example, surgery) or reduces the use of other interventions that are more expensive in the United States, then the system savings associated with pharmaceuticals are also higher. Given the other prices in the system of US healthcare,

Figure 2
Pharmaceutical and Healthcare Spending, Selected Countries



Source: OECD Data Explorer.
Note: Spending on prescription drugs is reported in the OECD Data Explorer as expenditures on "Pharmaceuticals and other medical non-durable goods." Per capita spending is in PPP-adjusted constant US dollars.

perhaps drug prices are not exceptionally high. Put another way, the market for pharmaceuticals may not be more problematic than other inputs, and interventions in the market for one input may introduce unexpected distortions.

Information asymmetries and other market imperfections can result in distorted prices, and therefore research and development incentives. Intervention by governments can, in theory, address these problems. But the fundamental challenge of valuing new treatments remains. To payers, the costs are visible and

immediate; the benefits may accrue over a long period of time. Government payers may introduce distortions by breaking the link between social value and profits. Prior research has identified distortions resulting from Medicaid drug pricing policies; for example, see Duggan and Scott Morton (2006) and Alpert, Duggan, and Hellerstein (2013). This is particularly a risk when budgets are strained by high demand for a novel product: it may be difficult to distinguish between products with high demand because they yield therapeutically important results from those with high demand due to moral hazard or well-organized patient advocates. Denying coverage can also be politically fraught. After the National Institute for Clinical Excellence recommended against reimbursement of several cancer drugs, the United Kingdom established the Cancer Drug Fund to pay for them in response to patient outcry and legal challenges (Aggarwal, Ginsburg, and Fojo 2014). Subsequent analyses generally conclude that this funding did not provide value to UK cancer patients (Aggarwal et al. 2017). Similar political pressures may influence Medicare decisions in the United States, where warnings about "death panels"[3]—committees of bureaucrats who determine whether patients will have access to care—have long resonated.

Generic drug competition in the United States is generally robust, with rapid entry and price declines following patent expiration. This is exactly what economic theory predicts: high profits are competed away when barriers to entry are reduced and products are undifferentiated. Margins on many generic products are very small (so small that manufacturer exit has created shortages, particularly in some markets for injectable products). Indeed, as noted in the introduction, generic prices in the United States are lower than in peer countries. The high prices paid by US consumers during the limited patent term are at least somewhat offset by the benefits of low prices later on.

The General Accounting Office estimated that the average profit margin of the largest drug companies was 15–20 percent from 2006 to 2015, versus 4–9 percent across the largest 500 firms (Government Accountability Office 2017). The Congressional Budget Office reported that US-based publicly traded pharmaceutical firms spent an average of more than 19 percent of net revenues on research and development between 2000 and 2019, much higher than the average of 3 percent for the S&P Total Market Index and even above other innovative sectors like software and semiconductors (Congressional Budget Office 2021). Industry lobbyists sometimes justify drug prices by pointing to these large research and development expenses. Those investments are made with the expectation of future profits; they should not be a justification after the fact. However, policies that lower expected profits reduce research and development spending, and by extension, innovation (Acemoglu and Linn 2004; Filson 2012; Blume-Kohout and Sood 2013; Dubois et al. 2015). The size and importance of the US market make such policy changes much more significant than in most other countries. We have little evidence on the value of the marginal research and development that is sacrificed, though Dranove, Garthwaite,

[3] Sarah Palin, former Republican Vice Presidential candidate, first used this term during the debate over the Affordable Care Act.

and Hermosilla (2022) and Finkelstein (2004) suggest that increases to expected market size induced by policy changes in the United States did not trigger scientifically novel innovation. Policies that target the rewards to more innovative products through longer exclusivity, such as for orphan drugs, or through faster regulatory approval, such as the US Food and Drug Administration's breakthrough therapy designation, appear to have been more successful in this regard (Yin 2008; Chandra et al. 2024).

Firms themselves may have difficulty identifying which products are likely to be the most profitable. Not only do they face technical risk during the process of drug development, but they also face market risk if a product completes clinical trials. Wouters, McKee, and Luyten (2020) find no correlation between the research and development costs associated with specific products and their prices. Reports by consulting firms find that new drug launches often fail to meet analysts' expectations (McKinsey & Company 2013; Deloitte Center for Health Solutions 2020), with poor formulary placement or formulary exclusions as an explanation in many cases. Greater clarity from payers could reduce this uncertainty, and allow firms to direct their efforts more efficiently.

To achieve better outcomes than in the current system, policymakers must focus on more than reducing the average price paid or total spending. Accepting higher prices on effective drugs (offset by reducing the rewards to ineffective therapies), and valuing long-run benefits accrued over time, should result in more efficient use and innovation incentives. Unlike many other health innovations, pharmaceuticals have low transportation costs, are easy to adopt and deliver, and see large price reductions within 10–15 years of introduction. Their potential in treating diseases like Alzheimer's, obesity, and others justifies at least maintaining, if not increasing, incentives for investment. The informational and political challenges are nontrivial, however. Addressing some of the frictions or imperfections described here should also increase efficiency.

References

Association of the British Pharmaceutical Industry (ABPI). 2014. *Understanding the 2014 Pharmaceutical Price Regulation Scheme.* https://www.abpi.org.uk/media/544dgo5x/understanding-the-2014-pprs.pdf.

Acemoglu, Daron, and Joshua Linn. 2004. "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry." *Quarterly Journal of Economics* 119 (3): 1049–90.

Aggarwal, A., T. Fojo, C. Chamberlain, C. Davis, and R. Sullivan. 2017. "Do Patient Access Schemes for High-Cost Cancer Drugs Deliver Value to Society? Lessons from the NHS Cancer Drugs Fund." *Annals of Oncology* 28 (8): 1738–50.

Aggarwal, Ajay, Ophira Ginsburg, and Tito Fojo. 2014. "Cancer Economics, Policy and Politics: What

Informs the Debate? Perspectives from the EU, Canada and US." *Journal of Cancer Policy* 2 (1): 1–11.

Alpert, Abby, Mark Duggan, and Judith K. Hellerstein. 2013. "Perverse Reverse Price Competition: Average Wholesale Prices and Medicaid Pharmaceutical Spending." *Journal of Public Economics* 108: 44–62.

Amin, Krutika, Imani Telesford, Rakesh Singh, and Cynthia Cox. 2023. "How Do Prices of Drugs for Weight Loss in the US Compare to Peer Nations' Prices?" KFF, August 17. https://www.kff.org/health-costs/issue-brief/how-do-prices-of-drugs-for-weight-loss-in-the-u-s-compare-to-peer-nations-prices/.

Antonanzas, Fernando, Carmelo Júarez-Castelló, Reyes Lorente, and Roberto Rodríguez-Ibeas. 2019. "The Use of Risk-Sharing Contracts in Healthcare: Theoretical and Empirical Assessments." *PharmacoEconomics* 37 (12): 1469–83.

Avalere. 2023. "Survey Finds 58% of Payers Use Outcomes-Based Contracts." https://avalere.com/insights/58-of-payers-use-outcomes-based-contracts.

Azoulay, Pierre. 2002. "Do Pharmaceutical Sales Respond to Scientific Evidence?" *Journal of Economics and Management Strategy* 11 (4): 551–94.

Barros, Pedro Pita. 2011. "The Simple Economics of Risk-Sharing Agreements between the NHS and the Pharmaceutical Industry." *Health Economics* 20 (4): 461–70.

Batty, Michael, and Benedic Ippolito. 2017. "Financial Incentives, Hospital Care, and Health Outcomes: Evidence from Fair Pricing Laws." American Economic Journal: Economic Policy 9 (2): 28–56.

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

Boggild, Mike, Jackie Palace, Pelham Barton, Yoav Ben-Shlomo, Thomas Bregenzer, Charles Dobson, and Richard Gray. 2009. "Multiple Sclerosis Risk Sharing Scheme: Two Year Results of Clinical Cohort Study with Historical Comparator." *BMJ* 339: b4677.

Borrell, Joan-Ramon. 1999. "Pharmaceutical Price Regulation: A Study on the Impact of the Rate-of-Return Regulation in the UK." *PharmacoEconomics* 15 (3): 291–303.

Boyle, Jemma M., Gemma Hegarty, Christopher Frampton, Elizabeth Harvey-Jones, Joanna Dodkins, Katharina Beyer, Gincy George, Richard Sullivan, Christopher Booth, and Ajay Aggarwal. 2021. "Real-World Outcomes Associated with New Cancer Medicines Approved by the Food and Drug Administration and European Medicines Agency: A Retrospective Cohort Study." *European Journal of Cancer* 155: 136–44.

Brekke, Kurt R., Tor Helge Holmas, and Odd Rune Straume. 2011. "Reference Pricing, Competition, and Pharmaceutical Expenditures: Theory and Evidence from a Natural Experiment." *Journal of Public Economics* 95 (7): 624–38.

Carlson, Josh J., Shuxian Chen, and Louis P. Garrison Jr. 2017. "Performance-Based Risk-Sharing Arrangements: An Updated International Review." *PharmacoEconomics* 35 (10): 1063–72.

Centers for Medicare and Medicaid Services. 2021. "Most Favored Nation Model." https://www.cms.gov/priorities/innovation/innovation-models/most-favored-nation-model.

Chandra, Amitabh, Jennifer Kao, Kathleen L. Miller, and Ariel D. Stern. 2024. "Regulatory Incentives for Innovation: The FDA's Breakthrough Therapy Designation." *Review of Economics and Statistics*. https://doi.org/10.1162/rest_a_01434.

Chatterjee, Arnaub, Casey Dougan, B. J. Tevelow, and Amir Zamani. 2017. *Innovative Pharma Contracts: When Do Value-Based Arrangements Work?* McKinsey and Company.

Ching, Andrew T., and Masakazu Ishihara. 2012. "Measuring the Informative and Persuasive Roles of Detailing on Prescribing Decisions." *Management Science* 58 (7): 1374–87.

Collins, Ben. 2020. "Access to New Medicines in the English NHS." The King's Fund, October 28. https://www.kingsfund.org.uk/insight-and-analysis/long-reads/access-new-medicines-english-nhs.

Commonwealth Fund. 2020. "Country Profiles." https://www.commonwealthfund.org/international-health-policy-center/countries.

Congressional Budget Office. 2021. *Research and Development in the Pharmaceutical Industry*. Congressional Budget Office.

Conti, Rena M., Richard G. Frank, and Jonathan Gruber. 2021. "Regulating Drug Prices While Increasing Innovation." *New England Journal of Medicine* 385 (21): 1921–23.

Costa, Eduardo, and Carolina Santos. 2022. "Pharmaceutical Pricing Dynamics in an Internal Reference Pricing System: Evidence from Changing Drugs' Reimbursements." *European Journal of Health Economics* 23 (9): 1497–1518.

D'Souza, Arielle, Kendall Hoyt, Christopher M. Snyder, and Alec Stapp. 2024. "Can Operation Warp

Speed Serve as a Model for Accelerating Innovations beyond COVID Vaccines?" NBER Working Paper 32831.

Dafny, Leemore S. 2005. "How Do Hospitals Respond to Price Changes?" *American Economic Review* 95 (5): 1525–47.

Danzon, Patricia M., and Li-Wei Chao. 2000. "Does Regulation Drive out Competition in Pharmaceutical Markets?" *Journal of Law and Economics* 43 (2): 311–58.

David, Guy, Sara Markowitz, and Seth Richards-Shubik. 2010. "The Effects of Pharmaceutical Marketing and Promotion on Adverse Drug Events and Regulation." *American Economic Journal: Economic Policy* 2 (4): 1–25.

Deese, Brian, Jonathan Gruber, and Ryan Cummings. 2024. "Ozempic Could Threaten the Federal Budget." *New York Times*, March 4. https://www.nytimes.com/2024/03/04/opinion/ozempic-wegovy-medicare-federal-budget.html.

Deloitte Center for Health Solutions. 2020. *Key Factors to Improve Drug Launches*. Deloitte LLP.

Department of Health and Human Services. 2023. "HHS and DOC Announce Plan to Review March-In Authority." Office of Public Affairs (press release), March 21. https://www.commerce.gov/news/press-releases/2023/03/hhs-and-doc-announce-plan-review-march-authority.

DiMasi, Joseph A., Henry G. Grabowski, and Ronald W. Hansen. 2016. "Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs." *Journal of Health Economics* 47: 20–33.

DiMasi, Joseph A., Ronald W. Hansen, and Henry G. Grabowski. 2003. "The Price of Innovation: New Estimates of Drug Development Costs." *Journal of Health Economics* 22 (2): 151–85.

Dranove, David, Craig Garthwaite, and Manuel Hermosilla. 2022. "Does Consumer Demand Pull Scientifically Novel Drug Innovation?" *RAND Journal of Economics* 53 (3): 590–638.

Dubois, Pierre, Olivier de Mouzon, Fiona Scott Morton, and Paul Seabright. 2015. "Market Size and Pharmaceutical Innovation." *RAND Journal of Economics* 46 (4): 844–71.

Duggan, Mark, and Fiona Scott Morton. 2006. "The Distortionary Effects of Government Procurement: Evidence from Medicaid Prescription Drug Purchasing." *Quarterly Journal of Economics* 121 (1): 1–30.

EFPIA. 2024. "EFPIA Patients W.A.I.T. Indicator 2023 Survey." https://www.efpia.eu/media/vtapbere/efpia-patient-wait-indicator-2024.pdf.

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.

Epstein, Andrew J., and Jonathan D. Ketcham. 2014. "Information Technology and Agency in Physicians' Prescribing Decisions." *RAND Journal of Economics* 45 (2): 422–48.

Fang, Hanming, and Qing Gong. 2017. "Detecting Potential Overbilling in Medicare Reimbursement via Hours Worked." *American Economic Review* 107 (2): 562–91.

FDA. 2019. "Trump Administration Takes Historic Steps to Lower US Prescription Drug Prices." FDA (press release), December 18. https://www.fda.gov/news-events/press-announcements/trump-administration-takes-historic-steps-lower-us-prescription-drug-prices.

Federal Trade Commission. 2022. "Federal Trade Commission Report on Pharmaceutical Product Hopping." https://www.ftc.gov/reports/federal-trade-commission-report-pharmaceutical-product-hopping.

Filson, Darren. 2012. "A Markov-Perfect Equilibrium Model of the Impacts of Price Controls on the Performance of the Pharmaceutical Industry." *RAND Journal of Economics* 43 (1): 110–38.

Finkelstein, Amy. 2004. "Static and Dynamic Effects of Health Policy: Evidence from the Vaccine Industry." *Quarterly Journal of Economics* 119 (2): 527–64.

Garrison, Louis P., Jr., Adrian Towse, Andrew Briggs, Gerard de Pouvourville, Jens Grueger, Penny E. Mohr, J. L. (Hans) Severens, Paolo Siviero, and Miguel Sleeper. 2013. "Performance-Based Risk-Sharing Arrangements—Good Practices for Design, Implementation, and Evaluation: Report of the ISPOR Good Practices for Performance-Based Risk-Sharing Arrangements Task Force." *Value in Health* 16 (5): 703–19.

Gavi, the Vaccine Alliance (GAVI). 2021. "Gavi Pneumococcal Conjugate Advance Market Commitment Pilot: 2nd Outcomes and Impact Evaluation." Gavi, the Vaccine Alliance.

Government Accountability Office. 2017. *Drug Industry: Profits, Research and Development Spending, and Merger and Acquisition Deals*. Government Accountability Office.

Grossman, Gene M., and Edwin L.-C. Lai. 2008. "Parallel Imports and Price Controls." *RAND Journal of Economics* 39 (2): 378–402.

Guardado, José R. 2024. "Competition in PBM Markets and Vertical Integration of Insurers with PBMs:

2024 Update." https://www.ama-assn.org/system/files/prp-pbm-shares-hhi-2024.pdf.

Haslam, Alyson, Jennifer Gill, Tyler Crain, Diana Herrera-Perez, Emerson Y. Chen, Talal Hilal, Myung S. Kim, and Vinay Prasad. 2021. "The Frequency of Medical Reversals in a Cross-Sectional Analysis of High-Impact Oncology Journals, 2009–2018." *BMC Cancer* 21 (1): 889.

Iizuka, Toshiaki. 2007. "Experts' Agency Problems: Evidence from the Prescription Drug Market in Japan." *RAND Journal of Economics* 38 (3): 844–62.

IQVIA Institute. 2024. *The Global Use of Medicines 2024: Outlook to 2028*. IQVIA Institute.

Iyengar, Swathi, Kiu Tay-Teo, Sabine Vogler, Peter Beyer, Stefan Wiktor, Kees de Joncheere, and Suzanne Hill. 2016. "Prices, Costs, and Affordability of New Medicines for Hepatitis C in 30 Countries: An Economic Analysis." *PLoS Med* 13 (5): e1002032.

Jacobson, Mireille, A. James O'Malley, Craig C. Earle, Juliana Pakes, Peter Gaccione, and Joseph P. Newhouse. 2006. "Does Reimbursement Influence Chemotherapy Treatment for Cancer Patients?" *Health Affairs* 25 (2): 437–43.

Kaiser, Ulrich, Susan J. Mendez, Thomas Rønde, and Hannes Ullrich. 2014. "Regulation of Pharmaceutical Prices: Evidence from a Reference Price Reform in Denmark." *Journal of Health Economics* 36: 174–87.

Kanavos, Panos, Anna-Maria Fontrier, Jennifer Gill, and Dionysis Kyriopoulos. 2017. *The Implementation of External Reference Pricing within and across Country Borders*. London School of Economics.

Kremer, Michael. 1998. "Patent Buyouts: A Mechanism for Encouraging Innovation." *Quarterly Journal of Economics* 113 (4): 1137–67.

Kremer, Michael, and Heidi Williams. 2010. "Incentivizing Innovation: Adding to the Tool Kit." *Innovation Policy and the Economy* 10: 1–17.

Kyle, Margaret K. 2007. "Pharmaceutical Price Controls and Entry Strategies." *Review of Economics and Statistics* 89 (1): 88–99.

Kyle, Margaret K. 2011. "Strategic Responses to Parallel Trade." *B. E. Journal of Economic Analysis and Policy* 11 (2).

Kyle, Margaret K. 2018. "Are Important Innovations Rewarded? Evidence from Pharmaceutical Markets." *Review of Industrial Organization* 53 (1): 211–34.

Kyle, Margaret K., and Anita M. McGahan. 2012. "Investments in Pharmaceuticals before and after TRIPS." *Review of Economics and Statistics* 94 (4): 1157–72.

Kyle, Margaret K., and Heidi Williams. 2017. "Is American Health Care Uniquely Inefficient? Evidence from Prescription Drugs." *American Economic Review* 107 (5): 486–90.

Leffler, Keith B. 1981. "Persuasion or Information? The Economics of Prescription Drug Advertising." *Journal of Law and Economics* 24 (1): 45– 74.

Maini, Luca, and Fabio Pammolli. 2023. "Reference Pricing as a Deterrent to Entry: Evidence from the European Pharmaceutical Market." *American Economic Journal: Microeconomics* 15 (2): 345–83.

McKinsey & Company. 2013. *Beyond the Storm: Launch Excellence in the New Normal*. McKinsey and Company.

Medscape. 2021. *Medscape Physician Compensation Report*. Medscape.

Milcent, Carine. 2021. "From Downcoding to Upcoding: DRG Based Payment in Hospitals." *International Journal of Health Economics and Management* 21 (1): 1–26.

Miraldo, Marisa. 2009. "Reference Pricing and Firms' Pricing Strategies." *Journal of Health Economics* 28 (1): 176–97.

Mooghali, Maryam, Aaron P. Mitchell, Joshua J. Skydel, Joseph S. Ross, Joshua D. Wallach, and Reshma Ramachandran. 2024. "Characterization of Accelerated Approval Status, Trial Endpoints and Results, and Recommendations in Guidelines for Oncology Drug Treatments from the National Comprehensive Cancer Network: Cross Sectional Study." *BMJ Medicine* 3 (1): e000802.

Mulcahy, Andrew W., Daniel Schwam, and Susan L. Lovejoy. 2024. "International Prescription Drug Price Comparisons: Estimates Using 2022 Data." *RAND Health Quarterly* 11 (3): 5.

National Institute for Health and Care Excellence. 2018. *Beta Interferons and Glatiramer Acetate for Treating Multiple Sclerosis*. National Institute for Health and Care Excellence.

Neumann, Peter J., James D. Chambers, Françoise Simon, and Lisa M. Meckley. 2011. "Risk-Sharing Arrangements That Link Payment for Drugs to Health Outcomes Are Proving Hard to Implement." Health Affairs 30 (12): 2329–37.

OECD. 2008. *Pharmaceutical Pricing Policies in a Global Market*. OECD.

Paris, Valérie, and Annalisa Belloni. 2013. "Value in Pharmaceutical Pricing." OECD Health Working Paper 63.

Patterson, Julie, James Motyka, and John Michael O'Brien. 2024. "Unintended Consequences of the Inflation Reduction Act: Clinical Development Toward Subsequent Indications." *American Journal of Managed Care* 30 (2): 82–86.

Piatkiewicz, Trevor Jozef, Janine Marie Traulsen, and Tove Holm-Larsen. 2018. "Risk-Sharing Agreements in the EU: A Systematic Review of Major Trends." *Pharmacoecon Open* 2 (2): 109–23.

Prasad, Vinay, Andrae Vandross, Caitlin Toomey, Michael Cheung, Jason Rho, Steven Quinn, Satish Jacob Chacko, et al. 2013. "A Decade of Reversal: An Analysis of 146 Contradicted Medical Practices." *Mayo Clinic Proceedings* 88 (8): 790–98.

Slaoui, Moncef, and Matthew Hepburn. 2020. "Developing Safe and Effective Covid Vaccines—Operation Warp Speed's Strategy and Approach." *New England Journal of Medicine* 383 (18): 1701–03.

Sood, Neeraj, Diane Ung, Anil Shankarand, and Brian L. Strom. 2018. "A Novel Strategy for Increasing Access to Treatment for Hepatitis C Virus Infection for Medicaid Beneficiaries." *Annals of Internal Medicine* 169 (2): 118–19.

Towse, Adrian, Louis Garrison, and Ruth Puig-Peiró. 2011. "The Use of Pay-for-Performance for Drugs: Can It Improve Incentives for Innovation?" In *Incentives for Research, Development, and Innovation in Pharmaceuticals*, edited by Walter García-Fontes, 69–80. Springer Healthcare.

Wouters, Olivier J., Martin McKee, and Jeroen Luyten. 2020. "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009–2018." *JAMA* 323 (9): 844–53.

Yin, Wesley. 2008. "Market Incentives and Pharmaceutical Innovation." *Journal of Health Economics* 27 (4): 1060–77.

Yu, Justin S., Lauren Chin, Jennifer Oh, and Jorge Farias. 2017. "Performance-Based Risk-Sharing Arrangements for Pharmaceutical Products in the United States: A Systematic Review." *Journal of Managed Care and Specialty Pharmacy* 23 (10): 1028–40.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 79–102

The Economics of Generic Drug Shortages: The Limits of Competition

Rena M. Conti and Marta E. Wosińska

The US market for "generic" off-patent prescription drugs is sometimes regarded as a textbook example of a competitive market. The marginal costs of making generic drugs are low, so when patents expire, generic drug manufacturers making copycat versions of the brand drive the price down significantly (Scott Morton and Kyle 2011; Ganapati and McKibbin 2023; Grabowski and Vernon 1992; Berndt and Aitken 2011). In 2023, 92 percent of US drug prescriptions were filled as generics, representing less than 13 percent of overall invoice spending on drugs (IQVIA 2024). The success of the generic drug market is touted as the marquee government policy promoting affordability and access to prescription drugs. The large savings generic competition provides helps government and commercial payers offset spending on high priced brands.

However, the US generic drug market appears to have become a victim of its own success (Hopp, Brown, and Shore 2022; Conti and Berndt 2020). Over 100 prescription drugs, largely generics, have been in shortage each year over the last decade and many of these shortages are persistent, lasting several years (IQVIA 2023; USP 2024). Recent generic drug shortages include the antibiotic amoxicillin, chemotherapy drugs carboplatin and cisplatin, medical fluids such as saline, emergency

■ *Rena M. Conti is Associate Professor Markets, Public Policy, and Law, Questrom School of Business, co-Director of the Technology Policy and Research Institute, and faculty affiliate of the Institute for Business Markets and Society all at Boston University, Boston, Massachusetts. She is also an elected member of the Conference on Research in Income and Wealth. Marta E. Wosińska is Senior Fellow, Center on Health Policy, The Brookings Institution, Washington, DC. Wosińska is the corresponding author at mwosinska@brookings.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241420.

drugs such as atropine and epinephrine, and medications to treat attention deficit hyperactivity disorder (FDA 2024).

Such drug shortages can undermine patient care, endangering patient health and lives through delays in treatment, rationing, potential substitution to less efficacious treatments, and increased risk of medication errors (Vail et al. 2017; Tucker et al. 2020; He et al. forthcoming). Shortages also cost patients time and often create significant anxiety (Huff 2023).

Shortages present an economic puzzle—if generic drug markets were competitive, prices would rise in the face of a shortage, encouraging entry and reducing the number and persistence of shortages. Also, if shortages cause so much patient harm, then why do markets not recognize the value of supply chain reliability by paying a premium for it?

We argue that these puzzles can be explained by the inability of prices to adjust easily, the presence of asymmetric information, and agency problems coexisting in the US market for generic drugs. Prices for generic drugs do not rise easily because of contracts and government regulations. Entry costs can be substantive in part due to government-mandated manufacturing standards needed to assure drug quality manufacturing, safety, and efficacy. Those regulations are important because prescription drugs have strong "credence good" characteristics—that is, buyers cannot assess quality, safety, or efficacy and therefore the government imposes requirements on manufacturers (Conti, Frank, and Cutler 2024). Agency problems arise because buyers, including hospitals and retail pharmacies, and the payers, including commercial and government health plans, undervalue and underpay for supply chain reliability.

We begin by describing the scope of the generic drug shortage problem and a framework for how shortages arise and why they do not resolve easily. We follow with a discussion of the economic roles and incentives that are faced by key market participants. We conclude by discussing the current policy landscape.

Scope of the Prescription Drug Shortage Problem

For economists, a "shortage" typically refers to a situation where quantity demanded exceeds quantity supplied because price is artificially held below its equilibrium level. In the context of prescription drugs in the US market, there are two definitions that are used by US government and stakeholders. Neither definition explicitly accounts for price, instead focusing on whether inventory levels return to a baseline either for a specific version or all the versions combined. Here we discuss the definitions, levels, trends, and correlates of drug shortages.

What Is a Drug "Shortage"?

Two different measures of drug shortages are commonly cited in the press, government reports, and academic research (Wosińska, Fox, and Jensen 2015; Hopp, Brown, and Shore 2022).

One common drug shortage metric, from the US Food and Drug Administration (FDA), considers whether the combined supply from all manufacturers in a specific drug market meets historical demand patterns, which may include historical utilization levels or historical inventory levels. FDA defines a drug in shortage at the ingredient-route level (for example, the injectable antibiotic doxycycline). FDA will not post a shortage if manufacturers provide evidence that they have sufficient inventory to cover the existing shortfall or if they are ramping up production and are meeting all demand (Wosińska, Fox, and Jensen 2015). FDA uses similar logic when determining whether to take a drug off its shortage list.

The other commonly used metric, from the American Society for Health Pharmacists (ASHP), defines a drug shortage as a supply disruption that affects how the pharmacy prepares or dispenses a drug product or that influences patient care (Fox, Sweet, and Jensen 2014). The ASHP defines a drug in shortage at the specific ingredient-route-presentation level. Presentation might include different bag sizes (say, 500 versus 250 milliliters) of the same drug or whether the product comes in a prefilled syringe or must be drawn from a vial. The ASHP does not have the same visibility into the supply chain as the Food and Drug Administration, so it relies more heavily on reports from healthcare providers for reporting and delisting shortages.

The difference in shortage definitions has significant implications for how drug shortages are identified and counted. The aggregation-level differences mean, for example, that a single Food and Drug Administration shortage of a 5 percent dextrose injection may be counted as three 5 percent dextrose shortages in the American Society for Health Pharmacists database, differentiated by bag size. FDA's market-wide assessment also means that FDA may not list a shortage of a drug with alternate package sizes available. Additionally, the ASHP may keep shortages on their list long after FDA no longer considers them market-wide shortages.

Patterns in Drug Shortages

Given the differences in shortages definitions, it is not surprising that drug shortage counts and trends differ between the two sources. In 2024, the Food and Drug Administration drug shortage count was at over 120 drug shortages, much lower than the 2011 peak of 251 (FDA 2024). At the same time, in April 2024, American Society for Health Pharmacists announced that the number of shortages they track hit an all-time high of 323 (ASHP 2024).

Both data sources find that generics are more likely to be represented among shortage drugs than brand-name drugs. Between 2018 and 2023, 71 percent of drug shortages listed by the Food and Drug Administration were generic drugs (Eastern Research Group 2025). Of the 323 current American Society for Health Pharmacists shortages, 89 percent are generics, of which 12 percent are sole-source generics (ASHP 2024).

Analyses of Food and Drug Administration data suggest that price is related to shortages. Most drugs were listed on FDA's drug shortage database invoice at the pharmacy level for less than $5; 66 percent of solid oral medicines in shortage invoice for less than $3 while over half of sterile injectables in shortage were under

$5 (USP 2024). A recent analysis looked at the relative likelihood of shortage, finding that while only 1 percent of drugs invoiced at $500 or more are in shortage, 11 percent of drugs priced under $1 are in shortage (IQVIA 2023).

Both databases show a strong presence of sterile injectables, which are physician-administered drugs such as cancer therapies, intravenous nutrition, intravenous antibiotics, crash cart drugs to revive trauma patients, and morphine. The Food and Drug Administration's data suggest the rate of injectables to hover between 61 and 74 percent over the last ten years (USP 2024), while American Society for Health Pharmacists data pin the injectables share at 42–73 percent of shortages depending on the year (ASHP 2024).

Both databases show that shortages are persistent, and that persistence is increasing. IQVIA analysis of FDA drug shortages finds that 75 percent have been active for more than a year and 58 percent have been ongoing for more than two years. Almost one-quarter (27 products) have been in shortage for more than five years (USP 2024). One full year was added to average shortage time since 2020, when the average shortage spanned two years (USP 2024). Among the 323 current ASHP shortages, 24 percent began in 2020 or earlier (ASHP 2024).

Finally, generic drug shortages appear more prevalent in the United States than in other countries (Mulcahy et al. 2021). Only Canada reports drug shortages coincident with United States, likely because of common supply chains. Perhaps counterintuitively, a recent analysis finds that multi-source generic molecules are more likely to be in shortage than sole-source generic molecules (IQVIA 2023).

Anatomy of a Drug Shortage

Individual drug shortages come about for specific reasons, but the overall structure is similar: a shortage happens when the relevant drug supply chain cannot quickly adjust to an abrupt change in demand or supply.

We follow the published literature and term abrupt changes in demand or supply a "shock." In this section, we describe types of shock that may trigger drug shortages, factors that may enhance the size of the shock, and buffers that could make supply chains more likely to withstand shocks (Hopp, Brown, and Shore 2022; Wosińska, Mattingly, and Conti 2023). We also describe factors that drive the speed of recovery from a shortage and measures that the US government and supply chain participants take to minimize patient harm during shortages.

Types of Demand and Supply Chain Shocks

Demand shocks for drugs can occur for many reasons. A familiar one is an increase in disease prevalence (Hopp, Brown, and Shore 2022; Frank, McGuire, and Nason 2021). For example, COVID-19 rapidly increased demand for ventilator drugs, and the post-pandemic rise in respiratory diseases drove demand for amoxicillin. Chemical, biological, radiological, and nuclear threats for which the government prepares could cause major demand increases for medical

countermeasures (ASPR 2022; ASPR 2023). A demand shock can also be a drastic change in how a drug is used, as has been the case with the increase in the use of GLP-1 inhibitors for weight-loss. Additionally, demand can also spill over from a drug in shortage to another drug that might serve as substitute. For example, shortages in the anticoagulant heparin drove up hospital demand for an alternative drug substitute enoxaparin (Park, Carson, and Conti 2023).

Supply shocks can also occur for many reasons: manufacturing quality problems, natural disasters, manufacturers discontinuing select products in their portfolio, and even disruptions in international trade due to geopolitical conflicts (FDA 2019b; Wosińska, Mattingly, and Conti 2023; Hopp, Brown, and Shore 2022; ASPR 2023). These disruptions can occur at any stage of the production process, from raw materials to production of active pharmaceutical ingredients, inactive but critical ingredients, the finished dosage form of a drug, and delivery mechanisms such as syringes.

Manufacturing quality problems—lapses in manufacturing practices that require full or partial shutdowns of facilities or lines—have persistently topped the list of reasons for drug shortages: 56 percent of the total in 2011, 62 percent between 2013 and 2017, and 46 percent in 2022 (Woodcock and Wosińska 2013; FDA 2019a; Corrigan-Curay 2024). These issues relate to equipment operation and maintenance, quality of materials, process control, quality assurance, and thorough investigations of any manufacturing problems that arise (Woodcock and Wosińska 2013). To address such issues, often uncovered during inspections by the Food and Drug Administration, manufacturers may temporarily shut down or slow down production (Stomberg 2016), leading to a potentially dramatic drop in output.

Natural disasters, product discontinuations, and demand increases have also been associated with shortage reports over the years. To the extent there is a pattern to these attributions, it is the increased frequency of demand-driven shocks beginning with the COVID-19 pandemic. Drug shortages attributed to demand increases reached an all-time high in 2022 of 29 percent (FDA 2024).

No shortages in the last 20 years have been attributed to export restrictions related to geopolitics (Wosińska 2024), but concerns around potential export restrictions rose during the COVID-19 epidemic when the United Kingdom banned the export of over 80 drugs, and India restricted the export of over two dozen active pharmaceutical ingredients and finished drug products (Rees 2020a; Rees 2020b). Concerns around the exposure of supply chains to China have also caught Congressional attention (FDA 2019a).

There is very limited data available on the location of drug manufacturing, or its volume, which limits the ability of consumers, producers, and policymakers to evaluate exposure to geopolitical risks—including drug supply from countries with which the United States has an adversarial relationship. While the United States continues to be the primary location for "finished dosage form" production sites (41 percent in 2019), the share of "active pharmaceutical ingredient" sites outside the United States now stands at about 87 percent (Conti and Berndt 2020). Just before the pandemic, China represented 8 percent of finished dosage form

production sites and 18 percent of active pharmaceutical ingredients sites. Exposure to drugs produced in China is higher in antibiotics and heparin (Conti and Berndt 2020). It is also higher for key starting materials (Wosińska 2024).

Factors Affecting Relative Size of Shocks

Whether a shock causes a shortage depends, to some extent, on how large the shock is relative to market size. Here we discuss the extent to which the relative size of the shock is determined by how the market is structured or operates (Hopp, Brown, and Shore 2022; Wosińska, Mattingly, and Conti 2023).

One factor that affects relative shock size is concentration of production. The generic market, especially the sterile injectable market, is dominated by monopoly- and duopoly-supplied drugs. The extent of competition is greatest for the oldest generic drugs and is smallest for the most recent generic entrants. Firm revenues per generic drug are largest for young drugs and are heavily right skewed (Conti and Berndt 2020). Generic injectable markets experience less entry and higher rates of exit (Frank, Hicks, and Berndt 2021).

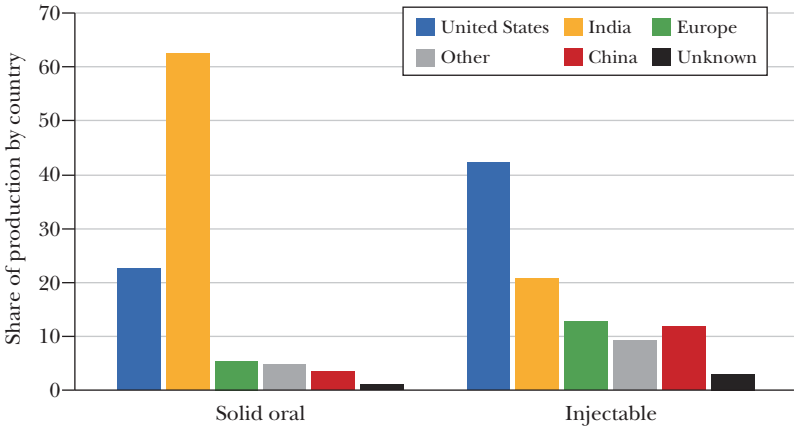
This dynamic is partly driven by mergers and acquisitions that have altered the scale and scope of supply for generic sterile injectables in the United States (IQVIA 2023; Conti and Berndt 2020; Berndt, Conti, and Murphy 2017; Kim and Scott Morton 2015; Sivashanker, Fanikos, and Kachalia 2018; Frank, McGuire, and Nason 2021). Several major sterile injectable facilities closed recently in the United States (US Senate Committee on Homeland Security and Governmental Affairs 2023), and major US players are further reconsidering their generic portfolios (Becker 2023b).

Many large generic drug firms have multiple facilities. Despite that, for economic reasons, they tend to concentrate production, instead of spreading it across facilities they own (Woodcock and Wosińska 2013). As a result, a process disruption on individual production lines can turn into market-wide drug shortages for a whole group of products (Woodcock and Wosińska 2013; Hopp, Brown, and Shore 2022). For example, after one key facility with production lines dedicated to cancer drugs shut down in early 2010, the number of shortages of chemotherapy drugs increased in short order from 4 to 24 (Woodcock and Wosińska 2013; McBride et al. 2013).

Another market structure factor that may create disproportionately large shocks is co-location of facilities. For example, tax policies implemented in the 1950s encouraged significant expansion of the US pharmaceutical industry in Puerto Rico (Bomey 2017; Jarvis 2018). When Hurricane Maria hit the island in September 2017, most if not all the 50 pharmaceutical facilities on the island were affected (Berndt, Conti, and Murphy 2018). Similarly, the low cost of labor and capital in East Asia, coupled with government subsidies in some of these countries, has shifted much of pharmaceutical production to India and China, creating a potential geopolitical threat that could span drugs across the spectrum, not just the currently vulnerable drugs like generic sterile injectables (FDA 2019a; Wosińska 2024).

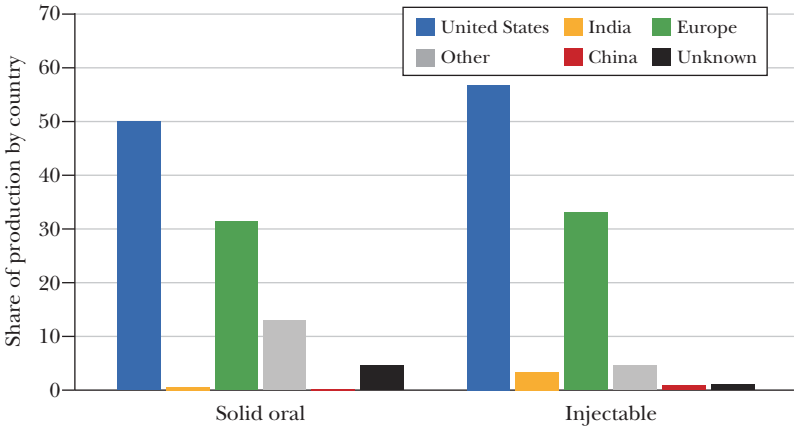
Geopolitical exposure is related to location. Figure 1 shows how much the geographic distribution differs across different generic drug types. Indian factories handle an overwhelming share of solid oral dose generics, but the United States and

Figure 1
Share of Generic Drug Production Volume by Location (2024)



Source: Wosińska (2025).
Note: Solid oral dose volume is in extended units (like number of tablets); injectable volume is in "eaches" (for example, number of vials or intravenous bags).

Figure 2
Share of Branded Drug Production Volume by Location (2024)



Source: Wosińska (2025).
Note: Solid oral dose volume is in extended units (like number of tablets); injectable volume is in "eaches" (for example, number of vials or intravenous bags).

Europe continue to lead in the production of generic sterile injectables. Figure 2 shows the geographic distribution of branded drug manufacturing. The contrast with branded drugs is striking, but not surprising—India's role is significantly smaller and China's is nonexistent.

Shock Buffers

Even if the shock size is relatively large, a shortage need not result if sufficient buffers exist to absorb the shock. Here we discuss two such buffers, inventory and spare capacity, explaining how they help buffer against shocks and the extent to which they are deployed.

How much inventory hospital and retail pharmacies carry depends on product turnover, demand variability, storage costs, and the consequences of not having the product available. Generally, just-in-time inventory management rules (Fox et al. 2009). Pharmacies may carry higher inventory for products with greater demand variability, if the drugs have to be administered immediately. But pharmacies may not carry certain products at all if treatment can be scheduled in advance or a day's delay in dispensing is acceptable. Additionally, hospitals generally carry lower inventories of products that have high holding costs, either because of their price point (like branded drugs) or their storage costs (like saline).

There have been proposals to encourage hospitals to carry higher inventory levels for critical drugs (CMS 2023; CMS 2024), but holding inventory downstream as a buffer is inefficient, partly because misallocation of product is difficult to correct once the product is at the point of care. Also, wholesalers and manufacturers are under pressure to optimize their inventory for maximum profit. Wholesalers keep the minimum on hand to respond to typical fluctuations in demand (Hopp, Brown, and Shore 2022). Drug manufacturers also use just-in-time inventory management and receive raw materials or active pharmaceutical ingredients only as they need them, improving production efficiency and reducing inventories (Hopp, Brown, and Shore 2022).

It is important, moreover, to distinguish between hospitals carrying inventory in anticipation of potential shocks and purchasing large quantities after a signal of a supply chain vulnerability, such as news of a hurricane or product quality issues at a facility. As we discuss above, the former is not happening enough, however the latter happens frequently, enabled by liberal return policies that allow hospitals to return drugs until close to expiration date (McKesson n.d.). Wholesalers can be slow in putting drugs on allocation, so quick-moving hospital systems can stockpile much larger quantities than they could ever use, lowering product availability for other, usually smaller hospitals or hospitals that need it most. For example, hospitals stockpiled prescription drugs that could help manage COVID-19 patients—sales volume of those drugs rapidly increased in the first two months of the pandemic—before most hospitals experienced an increase in demand (Park et al. 2023a).

Spare manufacturing capacity is another important buffer because it allows for expansion in production before inventories run out. However, with high overhead, there is little economic incentive for manufacturers to maintain spare capacity. Therefore, capacity generally must come from shifting production schedules around to accommodate expanded production of drugs in shortage.

Shifting a production lineup to accommodate increased demand for shortage products faces economic and technological challenges. On the economic front,

manufacturers may either have commitments on other products or can obtain better prices on other products than the product in shortage. On the technology front, production lines may not be appropriate for making drugs in shortage. In some cases that is because production lines are specialized (you cannot make one-liter IV bags on a line that only makes 250-milliliter IV bags). In other cases, it is because production lines are often dedicated to specific products (as with cancer drugs and antibiotics) as process requirements make switching lines either costly or impossible (Woodcock and Wosińska 2013).

Despite the lack of fungibility in production, most manufacturers of generic sterile injectables do not arrange backup facilities. One analysis of generic sterile-injectable products approved between 2000 and 2011 found that only just over 1 percent referenced more than one facility for production of the finished drug (Woodcock and Wosińska 2013). This contrasted with branded sterile-injectable applications, almost 20 percent of which were approved with backup facilities. About 5–6 percent of the studied generic sterile-injectables that submitted a generic drug version for approval have later submitted additional production sites, but none within one year of original submission of approval. There are no updated numbers showing the extent of backup facility use, but market dynamics suggest that increased use of such backups is unlikely.

To remedy the lack of private incentives, the US government has put in place programs to expand production capacity, but the expenditures have been limited and focused on drugs needed to deal with pandemics or chemical, biological, nuclear, or radiological events (Disbrow 2021). The federal government also keeps a Strategic National Stockpile that includes some of the drugs that are at risk of shortage (Hopp, Brown, and Shore 2022). But the drugs held in the stockpile are also in preparation for more extreme events, and the stockpile is only opened for national health emergencies, which do not cover ongoing numerous and persistent drug shortages.

Factors Affecting Recovery from a Shortage

In economics, resolution of a shortage would require that either supply meets demand or there is a price adjustment to reflect new demand and supply conditions. However, prices of generic drugs may not be able to easily adjust to new market conditions. Moreover, stakeholders, including the Food and Drug Administration, prefer supply to come back to pre-shortage levels, rather than after prices rise to reflect new conditions of scarcity.

There are several reasons why generic drug prices do not adjust easily when shortages occur. One reason has to do with contracts that hospitals, through their intermediaries, negotiate with generic manufacturers (Wosińska and Frank 2023). These contracts typically last one to three years and do not specify minimum quantity purchase amounts. Often these contracts include "best-price guarantees," meaning that if another party, not subject to a contract, offers a lower price, the contract holder must meet that lower price or the contract becomes invalid. Many contracts between commercial health plans and drug sellers may also contain best-price guarantees.

Prices for generic drugs are also constrained through government regulations. All drugs in the United States are subject to "most-favored-nation" contracts—contracts that require one or several payers to be given the lowest price offered to any other payer—related to government programs and private contracts that limit price increases (Berndt and Newhouse 2012; Conti et al. 2022). The Medicaid rebate program requires state-sponsored Medicaid insurance plans to purchase drugs at the "best price," entailing the payment of price rebates per unit sold (Duggan and Scott Morton 2006). Furthermore, the government has imposed inflation rebates on generic drug manufacturers, requiring them to pay rebates to the Medicaid program if the average price they charge to all buyers increases faster than the Consumer Price Index (Manning and Selck 2017).

In addition, over 40 percent of all nonprofit and public general acute-care hospitals participate in what is called the "340B program," created in 1992 to allow nonprofit hospitals that serve large numbers of low-income and uninsured to acquire drugs at deeply discounted prices. These discounts are directly tied to the Medicaid inflation rebates described above (Nikpay, Buntin, and Conti 2018). Furthermore, various state laws prohibiting "price gouging" act to limit price increases (Park, Carson, and Conti 2023).

Even if prices for generic drugs could adjust, technological constraints additionally limit how quickly or easily production could increase. Manufacturers generally face capacity constraints in the short-term, and they generally lack sufficient economic return to expand or free up capacity in the longer term. The costs and returns associated with stepping up production vary across products (for example, tablets versus injectables or large market drugs versus small batch drugs) and manufacturers (for example, those that are in the market versus not in the market, or whether a firm has higher value products on the line).

Capacity, especially for more complex products like generic sterile injectables, may be constrained by the lack of fungibility in the production process (Woodcock and Wosińska 2013; Hopp, Brown, and Shore 2022). Additionally, all changes to the approved production process require notification to the Food and Drug Administration, with larger changes requiring review (Center for Drug Evaluation and Research 2004). Such review can take several months, but FDA will expedite it in shortage situations.

Increasing production of one drug may mean lowering production of another. This could be worthwhile only if margins are attractive. Contracting out production could be an option for addressing capacity constraints. But aside from a Food and Drug Administration review of new production sites, contracts with third-party manufacturers demand a long-term price and quantity commitment, which is something few generic drug manufacturers are willing to make if future demand for the generic drug in shortage is uncertain.

The economics of creating new capacity is also challenging for the same uncertainty reasons. Building new capacity is expensive and can take several years, especially for sterile injectable fill and finish facilities (Scott Morton 1999; Wang, Li, and Anupindi 2023; Ganapati and McKibbin 2023). Even though Medicaid inflation

rebates do not apply to new entrants, there is little guarantee that any premium prices could be sustained before competition resumes.

For these reasons, resolution of many generic drug shortages often comes down to restoring previously disrupted production lines.

To help restore capacity, the US government can step in (HHS 2025). The 2024 response to the saline shortage is a great example. In that case, Hurricane Helene flooded a Baxter facility manufacturing 60 percent of United States supply of one-liter saline IV bags (Kansteiner 2024). While the Food and Drug Administration worked with the affected manufacturer to figure out a safe way to restore production quickly, other agencies stepped in to organize new supplies and to secure existing supplies. For example, the Department of Health and Human Services invoked the Defense Production Act to help Baxter obtain materials needed to clean and rebuild the affected facility, and the US Corp of Engineers built a bridge to ease access to the facility (HHS 2024).

Efforts to Minimize Patient Harm during Shortages

The US government and market participants have developed mechanisms for coping with drug shortages. Those efforts include allowing drug versions on the market that either do not meet review standards of the Food and Drug Administration or have not been reviewed by FDA for safety and efficacy. Those efforts also include allocating or reallocating product to patients most in need of treatment. We discuss these efforts here.

One way the Food and Drug Administration tries to mitigate the adverse impact of an impending or already present shortage is regulatory flexibility. This may include waiving its own quality requirements to allow distribution of medically necessary drugs in a shortage situation. For example, FDA allowed for distribution of one product with glass particles if accompanied by a letter instructing healthcare professionals to use a filter to remove the particles (Woodcock and Wosińska 2013). FDA may also allow importation of drugs that have not gone through the approval process in the United States, a flexibility some argue is underutilized (Bollyky et al. 2025).

Another regulatory band-aid that kicks in when a drug is already in shortage is pharmacy compounding—a practice where a pharmacist combines, mixes, or alters drug ingredients to create a medication without regulatory review from the Food and Drug Administration. Pharmacies can make variations of existing drugs (for example, make a liquid form of a drug for a pediatric patient) and are restricted from making a copy unless a drug is in shortage (Wosińska, Auchincloss, and Bernstein 2024). Their role in filling the gap for shortages of sterile injectable generics is limited by capacity and varies across products; for example, they are better equipped for compounding into vials or prefilled syringes than into IV bags.

When these interim measures are insufficient, there is a need to allocate product to patients who need it most. Within a hospital, staff may develop procedures for who should get the drug in shortage and who could wait until there is available supply or use an alternative treatment (Tucker et al. 2020). Allocation across hospitals and

providers is challenging for a different reason—no hospital or pharmacy wants to give up the inventory they may need themselves, so hospitals without product do not know to which of their peers they can turn for help. The US government has no visibility into where product sits nor power (nor probably willingness) to reallocate product between private entities. Allocation by wholesalers and manufacturers is largely limited to past purchase history, not current patient needs.

To fill the void in reallocation across hospitals, a nonprofit organization founded by a parent of a child affected by a cancer drug shortage does connect patients in need with small volumes of product withheld by wholesalers and manufacturers from traditional supply channels (Noguchi 2023). But scale and product availability through this program remain an issue.

Stakeholders and Their Roles in Generic Drug Shortages

In this section, we discuss the roles stakeholders play and the incentives they face or create with respect to generic drug supply chain reliability. We begin with the demand side (payers, pharmacies, wholesalers, hospitals, and group purchasing organizations) and then turn to the supply side (manufacturers and their regulators). We emphasize the role information asymmetry plays in the market for generic drugs.

Overview

For most generic drugs, patients receive prescriptions from physicians and fill them at retail pharmacies (like CVS, Walgreens, and Costco), through mail order pharmacies, or at specialty pharmacies (like CVS Caremark and Optum Rx). Typically, dispensed drugs come in solid-oral dose form, like tablets or capsules, but they can also be eye drops, creams, liquids, or device combinations such as autoinjectors or inhalers.

Other generic drugs, particularly sterile injectable generics, are administered by healthcare providers either in a hospital setting or an outpatient clinic. For these drugs, the provider carries the drug inventory, billing payers for products they use.

Virtually all generic prescriptions are covered and paid for by third party payers—that is, entities paying on behalf of patients. These include commercial health plans, employers, and public insurers like Medicare, Medicaid, the Veterans Affairs, the Military Health System in the Department of Defense, and the Indian Health System. Depending on their insurance and the setting (retail or hospital), patients may pay a share of the drug's cost.

Regulators include the Food and Drug Administration and, for certain controlled substances such as opioids, the Drug Enforcement Agency (DEA). FDA sets standard for all drug approvals and manufacturing, while DEA controls how much drugmakers can make of drugs they regulate. Manufacturers then respond to incentives set by regulators and payers, deciding, within the regulatory bounds, which drugs to make and how much to make.

Payers

Generally, payers incentivize physicians and pharmacies to select the least expensive version of generic drug available.

In the retail setting, payers reimburse pharmacies based on contracted rates that are either a share of list price or fixed reimbursement rate, plus a dispensing fee. In the inpatient setting, commercial and government payers often bundle inputs such as drugs as part of the overall cost of a certain a procedure, or within a hospital day or stay. In the outpatient clinic setting, lower-priced generics may also be bundled, or they are separately reimbursed using an average sales price across all the versions of the same generic. In all these cases, reimbursement is indifferent to variations in cost between generic versions, giving hospitals an incentive to seek the lowest-cost version.

These reimbursement mechanisms rest on the presumption that different versions of the same generic drug offered by competing manufacturers and the brand are therapeutically equivalent and therefore serve as perfect substitutes to each other. This presumption is not without merit—the Food and Drug Administration's approval of a generic means the manufacturer demonstrated to FDA that the product has the same efficacy and safety as the product they copy. However, this approach sidesteps the fact that reliability of manufacturing can be a critical differentiating factor across versions of the same drug.

Pharmacies and Wholesalers

State laws further encourage patients to use generics in the retail setting. Generic substitution laws, enacted in every state, either allow or require pharmacies to substitute lower-cost generics for higher-cost equivalent brands when available (Berndt and Newhouse 2012).

When dispensing generic drugs, retail and mail pharmacies are reimbursed not on the specific cost of the version they purchased, but on contracted rates that are either a share of list price or a fixed reimbursement rate (Berndt and Newhouse 2012). This in turn encourages pharmacies to purchase the lowest-cost version of a drug available.

To obtain best prices, major pharmacy chains participate in one of the four buying groups that represent over 90 percent of generic drug volume purchased (Fein 2021). These buying groups are joint ventures between wholesalers and major pharmacy chains (Fein 2018). For example, Red Oak Sourcing is a joint venture established between CVS Health and Cardinal Health in 2013. In 2014, Walgreens established the Walgreens Boots Alliance with Amerisource Bergen (now Cencora). In 2016, Walmart joined in with the wholesaler McKesson under the ClarusOne umbrella.

The rise of these buying groups, coupled with an increased capacity by the Food and Drug Administration to approve new generic applications (and therefore faster market entry), has driven significant deflation in the prices of many generic drugs (Fein 2021). This deflation may have adversely impacted generic drug manufacturer profitability.

Hospitals, Clinics, and Group Purchasing Organizations

Hospitals and clinics also face incentives from third-party payers to use the lowest-priced generic available. In the inpatient setting, commercial and government payers often bundle inputs such as drugs as part of the overall cost of a certain a procedure, or within a hospital day or stay. In the outpatient clinic setting, lower-priced generics may also be bundled, or they are separately reimbursed using an average sales price across all the versions of the same generic. In all these cases, reimbursement is indifferent to variations in cost between generic versions, giving hospitals an incentive to seek the lowest-cost version.

To obtain low prices for inputs, hospitals pool their bargaining power using group purchasing organizations (Burns 2022). The top three group purchasing organizations collectively represent hospitals that account for over 80 percent of hospital beds, giving them substantial collective purchasing power to negotiate prices. Vizient holds the largest share of the market with 37 percent of hospital beds, followed by Premier with 28 percent, and Health Trust with 15 percent.

When drug shortages occur, hospitals face various costs. One report found that hospitals incurred about $365 million in extra labor costs needed for shortage mitigation and $230 million in extra payments made to buy substitutes (Vizient 2019). Because switching from a drug in shortage to an alternative can often result in a different look across the versions of the same drug—for example, different formatting of the label or different size of the vial—the risk of medication errors can increase. Hospitals sometimes have had to send patients to other hospitals because they did not have a drug available (Hantel et al. 2019). Surveys consistently report providers and pharmacy staff experience stress resulting from shortages (Tucker and Daskin 2022; Fox, Sweet, and Jensen 2014).

Yet the behavior of hospitals suggests that they do not fully internalize the burden from drug shortages. There is little empirical data on price elasticity for hospitals, but anecdotal evidence suggests a large hospital system might switch a drug if faced with as little as $5,000 in annual savings (Wosińska and Frank 2023). This could mean switching a high utilization generic to another generic version because of a penny price difference, enabled by new data tools that readily identify savings opportunities (QuicksortRx 2023). There is also limited uptake of programs that help mitigate shortages, including programs that vet drug manufacturers for reliability and reward such manufacturers with longer-term contracts or set up buffer inventory (Wosińska and Frank 2023).

The low uptake of shortage mitigation programs may be linked to the limited financial consequences from shortages for hospitals, considering that $365 million and $230 million in extra expenses translates into 0.05 percent of the average budget of US hospitals (Wosińska and Frank 2023). It may also be reinforced by the fact that hospitals and clinics can and do charge "facility fees" on top of the costs of the services provided that may help defray cost increases from drug shortages (Howard et al. 2015; Conti and Berndt 2014). Furthermore, hospitals may face switching costs to such programs if their existing contracts have built-in disincentives for switching, as in the cases where a manufacturer bundles products

or the group purchasing organization has volume targets before a hospital can obtain rebates.

Regulators

Drugs have credence good characteristics because neither patients nor clinicians can discern quality manufacturing, safety, or efficacy. To assure that generic drugs meet safety and efficacy standards, they must first be approved by the Food and Drug Administration, using an abbreviated process (FDA 2011). The abbreviated pathway waves the need for certain clinical trials; however, it does not waive manufacturing approval requirements—all applicants must attest to FDA that the manufacturing processes they use is compliant with good manufacturing practices, and the product is safe and is similarly effective as the brand they copy.

After the product is on the market, the Food and Drug Administration oversees generic manufacturing quality—whether products are consistently made to specification—through periodic surveillance inspections, for-cause inspections following reports of potential manufacturing problems, and product testing (US FDA 2023; Berndt, Conti, and Murphy 2017).

The timing and methods of the Food and Drug Administration inspections after the product is on the market create a number of issues that reflect the classic enforcement dynamic: firms subject to oversight will be more likely to follow the rules if the probability and consequences of being caught are high (Becker 1968). Because foreign inspections are announced ahead of time, the effectiveness of the inspection program is weakened (GAO 2022). Between inspections, FDA relies on firms to file with a "defect report" on quality problems (FDA 2021), but filing a defect can lead to further scrutiny from FDA and drug manufactures may have a built-in disincentive to file such reports (Kaakeh et al. 2011).

The consequences of manufacturing quality problems can require costly remediation, but shortages also invite government support. Here is what might be referred to as a "too important to fail" problem: the Food and Drug Administration does not want shortages, so it will be flexible with its rules when problems arise, especially for medically necessary drugs where the manufacturer in violation of standards also holds a large market share (Woodcock and Wosińska 2013). But this sets up the wrong incentives, just like a traffic cop giving out warning tickets for speeding will have lesser deterrence impact than one giving out fines.

Manufacturers of controlled substances face added regulations through the Drug Enforcement Agency (GAO 2015). The finished dosage form of controlled substances products that are intended for US consumption must be manufactured in the United States. To manufacture drugs in the United States, manufacturers of both active ingredients and finished dosage form formulations not only require an FDA approval but also must be given a DEA production quota, which they cannot exceed. Manufacturers apply for these quotas and will be assigned some or all the amount they asked. DEA also imposes specific requirements on levels of inventory at various stages of the supply chain, along with restrictions on how this inventory be kept.

Manufacturers

Given the payer and buyer environment, manufacturers of generic drugs have a strong incentive to cut costs, potentially at the expense of reliability of product manufacturing.

One way cost-cutting has taken place is through significant offshoring that took off in the 1990s. A 2009 World Bank analysis suggested that wages in China were at the time 10 times lower and in India 12.5 times lower than for Western active ingredient manufacturers (Bumpas and Betsch 2009). A senior Food and Drug Administration official testified in Congress that "China has lower electricity, coal, and water costs. Chinese firms are also embedded in a network of raw materials and intermediary suppliers, and so have lower shipping and transaction costs for raw materials. They also face fewer environmental regulations on buying, handling, and disposing of toxic chemicals, leading to lower direct costs for these firms" (Woodcock 2019). An FDA (2011) report noted that despite potential productivity differences, manufacturing the active pharmaceutical ingredient of a drug in India could reduce costs for US and European companies by 30 to 40 percent.

Cost-cutting can show up in other ways. Facing low margins, with pressures to cut costs, firms lack incentives to upgrade facilities, may overuse their existing equipment, and may cut corners with respect to the tight manufacturing and quality control processes (Kansteiner 2023). Then, after being caught by the Food and Drug Administration for violations of quality standards, companies discard large batches of compromised product or recall such batches after releasing them to the market (Eglovitch 2022). To remedy persistent problems such as mold or metal shavings from machinery mixing in with product, manufacturers may have to shut down lines or facilities (Cundell 2016). In some cases, they close the sites because the cost to remedy problems is too high (Becker 2023a).

All these scenarios can lead to shortages in markets that are concentrated, where there are few buffers and where it can be challenging to restore production quickly.

Policy Issues and Open Questions

As we have described in this paper, the drug shortage problem is complex, but the economic causes of generic sterile injectable drug shortages in particular are well understood. Simply freeing prices to respond to demand and supply shocks will not address shortages because entry is costly and slow. Eliminating informational asymmetries by creating more transparency around product quality or product availability will also not be sufficient because stakeholders lack the sufficient incentive to act on such information. Hospitals and clinics are poor agents for patients in that they put too much weight on low prices in their drug purchasing decisions. As a result, the equilibrium manufacturing quality is below what is socially optimal, leading to shortages that are costly to patients. In this section, we highlight several areas of policy reform.

Reducing Government Policy Frictions

Ensuring the Food and Drug Administration has the staff to approve new applications for drugs with vulnerable supply and continue inspections to evaluate the quality of drug manufacturing are important twin regulatory goals. In addition, various researchers have pointed out that government policies, such as "most-favored-nation" pricing contracts and state price-gouging prohibitions, lower drug supply chain reliability by reducing price flexibility (Manning and Selck 2017; Augustine, Madhavan, and Nass 2018; Hopp, Brown, and Shore 2022). Others have pointed to Drug Enforcement Agency policies as culprits behind various controlled substance drug shortages (Wosińska 2024). What to do, if anything, about these policies is far from settled. Similarly, it is unclear what impact on supply chain reliability various tariff policies would have.

A better understanding of existing and proposed government policies requires resolving a more fundamental challenge: much of the data on drugs are proprietary or not in a structured format that makes empirical work easy (Hopp, Brown, and Shore 2022). For example, tariffs are generally applied to products based on where the active ingredient is made, not the final product, so figuring out which drug supply chains are exposed to potential tariffs is challenging (Wosińska 2025). Similarly, contract terms and prices, even when those prices are set by the US government, are not publicly available.

To address data challenges, researchers have had to scrape web archives to compile data from the Food and Drug Administration's website (Park et al. 2023; Stromberg 2016; Yurukoglu, Liebman, and Ridley 2017), request data through the Freedom of Information Act (Conti and Berndt 2020), or estimate supply, prices, and market shares from sources that are not publicly available (Galdin 2023). While improving data availability alone will not resolve shortages, finding ways to make more data available to stakeholders would be a good step forward in developing policies.

Paying for Quality

Multiple policy proposals have been put forward to address challenges related to manufacturing reliability of generics suppliers. Some proposals center around having someone, either the Food and Drug Administration or a "trusted third party," identify which manufacturers are more reliable and then have Medicare pay hospitals more if they purchase from those manufacturers (HHS 2024; Colvill et al. 2023; Dabestini et al. 2023). Other proposals set up incentives, either behavioral or outcomes-based, that motivate industry players to identify which manufacturers are reliable (Wosińska and Frank, 2023; US Senate Committee on Finance 2024; Wosińska and Frank 2024). A handful of proposals also contemplate potential penalties to hospitals that do not purchase reliably (HHS 2024; Schulman, Kumar, and Adashi 2023).

Advancing policy in this area will require spelling out how such proposals could be implemented to maximize impact and minimize unintended consequences.

Addressing Geopolitical Risks to Supply Chains

Reducing the exposure of the US drug supply chain to geopolitical adversaries such as China is another important policy objective (FDA 2019a). This concern has created a policy tension between addressing the causes of persistent shortages of generic injectable drugs and preventing potential shortages that could result from geopolitical conflicts—tension arising because the two problems require different solutions.

De-risking supply chains from China, or any other adversarial country, is no easy task given the enormity and complexity of supply chains that could potentially be compromised in a geopolitical conflict (Hopp, Brown, and Shore 2022). There are thousands of different drugs, each with a different set of inputs and many stages of production spanning the globe (Park et al. 2023). The cost of total de-risking would be staggering, and therefore any de-risking efforts must start with prioritization.

A starting point for prioritizing supply chains is a set of criteria for selecting which supply chains the US government would support and reinforce. For example, Hopp, Brown, and Shore (2022) argued policy interventions into drug markets and should be prioritized based upon at the expected patient harm from a shortage, the number of affected patients, and the probability of expected supply shortage in any given year. Any such prioritization framework should also account for common nodes in supply chains where a shortage of a single common ingredient could adversely impact availability of many different drugs (Wosińska and Frank 2022; Wosińska, Mattingly, and Conti 2023).

As it currently stands, operationalizing this framework is not easy. The existing lists of essential medicines do not necessarily reflect the list of drugs and drug components the absence of which would cause a public health crisis (Wosińska and Frank 2022; Wosińska, Mattingly, and Conti 2023). Worse yet, we have a poor understanding of the vulnerability of supply chains to various shocks, including geopolitical shocks (Park et al. 2023). Underpinning this lack of understanding is the paucity of data on where inactive ingredients and inputs in active ingredients are made. Much more research is needed to inform which supply chains are most vulnerable and to which kinds of shocks.

■ *Conti and Wosińska received grant funds from the National Science Foundation and Sloan Foundation to support past research efforts in this area. Wosińska received funding from West Health, the Commonwealth Fund, and Arnold Ventures to support her past research efforts in this area. Conti has received grant funds from the National Cancer Institutes, the Veterans Administration, the National Institute of Drug Abuse, the American Cancer Society, the Leukemia and Lymphoma Society, and Arnold Ventures unrelated to this manuscript. Conti has received consulting fees from Greylock McKinnon Associates, Analysis Group, and Keystone Consulting unrelated to this manuscript. Wosińska has received consulting fees from Greylock McKinnon Associates unrelated to this document. The authors thank Richard Frank, Erin Fox, Andy Wilson, Russell Findlay, Carlo di Notaristefani, Emily Tucker, Joey Mattingly, Elisabeth Reynolds, Robert Cook-Deegan, Scott Stern, Fiona Scott Morton, Ernst Berndt, and Jim Rebitzer for helpful discussions.*

References

ASHP. 2024. "Drug Shortages Statistics." https://www.ashp.org/drug-shortages/shortage-resources/drug-shortages-statistics.

ASPR (Administration for Strategic Preparedness and Response). 2022. *ASPR Strategic Plan for 2022–2026.* US Department of Health and Human Services.

ASPR. 2023. *National Health Security Environment and Threat Landscape National Health Security Strategy, 2023–2026.* US Department of Health and Human Services. Augustine, Norman R., Guru Madhavan, and Sharyl J. Nass. 2018. *Making Medicines Affordable: A National Imperative.* National Academies Press.

ASPR. n.d. "CBRN Resources." https://asprtracie.hhs.gov/cbrn-resources (accessed June 15, 2023).

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76 (2): 169–217.

Becker, Zoey. 2023a. "Bankrupt Akorn Pharma Calls It Quits and Closes All US Sites, Laying Off Entire Workforce." FiercePharma, February 23. https://www.fiercepharma.com/manufacturing/akorn-pharma-bankrupt-calls-it-quits-closes-all-us-sites-and-cuts-entire-workforce.

Becker, Zoey. 2023b. "In Strategic Pivot under New CEO Richard Francis, Teva Plots Cuts to Its Generics Portfolio." FiercePharma, May 23. https://www.fiercepharma.com/manufacturing/amidst-nationwide-shortages-teva-edit-its-generics-strategy-slash-loss-making.

Berndt, Ernst R., and Murray Aitken. 2011. "Brand Loyalty, Generic Entry and Price Competition in Pharmaceuticals in the Quarter Century after the 1984 Waxman-Hatch Legislation." *International Journal of the Economics of Business* 18 (2): 177–201.

Berndt, Ernst R., Rena M. Conti, and Stephen J. Murphy. 2017. "The Landscape of US Generic Prescription Drug Markets, 2004–2016." NBER Working Paper 23640.

Berndt, Ernst R., Rena M. Conti, and Stephen J. Murphy. 2018. "The Generic Drug User Fee Amendments: An Economic Perspective." *Journal of Law and the Biosciences* 5 (1): 103–41.

Berndt, Ernst R., and Joseph P. Newhouse. 2012. "Pricing and Reimbursement in US Pharmaceutical Markets." In *The Oxford Handbook of the Economics of the Biopharmaceutical Industry*, edited by Patricia M. Danzon and Sean Nicholson. Oxford University Press.

Bollyky, Thomas J. Sarosh N. Nagar, Chloe Searchinger, and Aaron S. Kesselheim. 2025. "The Role of Importation in Remediating U.S. Generic Drug Shortages." *New England Journal of Medicine* 392: 315–18.

Bomey, Nathan. 2017. "Hurricane Maria Halts Crucial Drug Manufacturing in Puerto Rico, May Spur

Shortages." *USA Today*, September 22. https://www.usatoday.com/story/money/2017/09/22/hurricane-maria-pharmaceutical-industry-puerto-rico/692752001/.

Bumpas, Janet, and Ekkehard Betsch. 2009. "Exploratory Study on Active Pharmaceutical Ingredient Manufacturing for Essential Medicines." World Bank HNP Discussion Paper 53075.

Burns, Lawton Robert. 2022. *The Healthcare Value Chain: Demystifying the Role of GPOs and PBMs*. Springer Nature.

Center for Drug Evaluation and Research. 2004. *Changes to an Approved NDA or ANDA*. Docket FDA-1999-D-0049. Food and Drug Administration.

CMS. (US Centers for Medicare and Medicaid Services). 2023. "Request for Public Comments on Potential Payment under the IPPS and OPPS for Establishing and Maintaining Access to Essential Medicines." *Federal Register* 88 (224): 587–91.

CMS. 2024. "CMS Proposes New Policies to Support Underserved Communities, Ease Drug Shortages, and Promote Patient Safety." CMS Newsroom, April 10. https://www.cms.gov/newsroom/press-releases/cms-proposes-new-policies-support-underserved-communities-ease-drug-shortages-and-promote-patient.

Colvill, Stephen, Thomas Roades, Gerrit Hamre, Marianne Hamilton Lopez, Cameron Joyce, Remi Shendell, and Mark McClellan. 2023. "Advancing Federal Coordination to Address Drug Shortages." Duke-Margolis Center for Health Policy. September 7. https://healthpolicy.duke.edu/publications/advancing-federal-coordination-address-drug-shortages.

Congressional Research Service. 2018. "Drug Shortages: Causes, FDA Authority, and Policy Options." Congressional Research Service in Focus, December 27. https://www.congress.gov/crs-product/IF11058.

Congressional Research Service. 2023. *The Strategic National Stockpile: Overview and Issues for Congress*. Congressional Research Service.

Conti, Rena M., and Ernst R. Berndt. 2014. "Specialty Drug Prices and Utilization after Loss of U.S. Patent Exclusivity, 2001–2007." NBER Working Paper, 20016.

Conti, Rena M., and Ernst R. Berndt. 2020. "Four Facts Concerning Competition in US Generic Prescription Drug Markets." *International Journal of the Economics of Business* 27 (1): 27–48.

Conti Rena M., Richard G. Frank, David M. Cutler. 2024. "The Myth of the Free Market for Pharmaceuticals." New England Journal of Medicine 390 (16): 1448–50.

Conti, Rena M., Brigham Frandsen, Michael L. Powell, and James B. Rebitzer. 2022. "Common Agent or Double Agent? Pharmacy Benefit Managers in the Prescription Drug Market." NBER Working Paper 28866.

Conti, Rena M., and Fiona Scott Morton. 2021. "Building a Resilient Rx Drug Supply: A New HHS Office and Other Steps." Health Affairs Forefront, August 27. https://www.healthaffairs.org/content/forefront/building-resilient-rx-drug-supply-new-hhs-office-and-other-steps.

Corrigan-Curay, Jacqueline. 2024. "State of Generics and Biosimilars 2024." Presentation, US Food and Drug Administration, October 23. https://subscriber.ipq.org/wp-content/uploads/2025/01/GRxBiosims-2024-PPT-Jacqueline-Corrigan-Curay.pdf.

Cundell, Tony. 2016. "Mold Monitoring and Control in Pharmaceutical Manufacturing Areas." *American Pharmaceutical Review*, July 30. https://www.americanpharmaceuticalreview.com/Featured-Articles/190686-Mold-Monitoring-and-Control-in-Pharmaceutical-Manufacturing-Areas/.

Dabestini, Arash, Sara Grossman, Joseph P. Nathan, Carl W. Bazil, Ryan C. Costantino, Erin R. Fox, Joe Graedon, et al. 2023. "A Data-Driven Quality-Score System for Rating Drug Products and Its Implications for the Healthcare Industry." *Journal of the American Pharmacists Association* 63 (2): 501–06.

Disbrow, Gary. 2021. "Testimony before the House Appropriations Subcommittee on Labor, Health and Human Services, Education, and Related Agencies." https://www.congress.gov/117/meeting/house/114232/witnesses/HHRG-117-AP07-Wstate-DisbrowG-20211117.pdf.

Duggan, Mark, and Fiona M. Scott Morton. 2006. "The Distortionary Effects of Government Procurement: Evidence from Medicaid Prescription Drug Purchasing." *Quarterly Journal of Economics* 121 (1): 1–30.

Eastern Research Group. 2025. *Analysis of Drug Shortages, 2018–2023*. US HHS.

Eglovitch, Joanne. 2022. "FDA Warns US Sterile Injectable Maker on Contamination Controls, Another US Maker Warned for Poor Building Conditions." Regulatory Affairs Professionals Society, October 28. https://www.raps.org/news-and-articles/news-articles/2022/10/fda-warns-us-sterile-injectable-maker-on-contamina.

FDA (US Food and Drug Adminstration). 2011. *Abbreviated New Drug Applications and 505(b)(2) Applications (Final Rule)*. Docket FDA-2011-N-0830. US Food and Drug Administration.

FDA. 2019a. "Safeguarding Pharmaceutical Supply Chains in a Global Economy." Congressional Testimony, October 30. https://www.fda.gov/news-events/congressional-testimony/safeguarding-pharmaceutical-supply-chains-global-economy-10302019.

FDA. 2019b. *Drug Shortages: Root Causes and Potential Solutions.* US Food and Drug Administration.

FDA. 2021. *Field Alert Report Submission Questions and Answers Guidance for Industry.* US Food and Drug Administration.

FDA. 2023. "Generic Drug User Fee Amendments." https://www.fda.gov/industry/fda-user-fee-programs/generic-drug-user-fee-amendments.

FDA. 2024. "FDA Drug Shortages. Current and Resolved Drug Shortages and Discontinuations Reported to FDA." https://www.accessdata.fda.gov/scripts/drugshortages/default.cfm.

Fein, Adam J. 2018. "Meet the Power Buyers Driving Generic Drug Deflation." Drug Channels. February 1. https://www.drugchannels.net/2018/02/meet-power-buyers-driving-generic-drug.html.

Fein, Adam J. 2021. "CVS Pharmacy Downsides: 10 Industry Trends Driving the Retail Shakeout." Drug Channels. December 1. https://www.drugchannels.net/2021/12/cvs-pharmacy-downsizes-10-industry.html.

Fox, Erin R., Annette Birt, Ken B. James, Heather Kokko, Sandra Salverson, and Donna L. Soflin. 2009. "ASHP Guidelines on Managing Drug Product Shortages in Hospitals and Health Systems." *American Journal of Health Systems Pharmacy* 66 (15): 1399–1406. https://doi.org/10.2146/ajhp090026.

Fox, Erin R., Burgunda V. Sweet, and Valerie Jensen. 2014. "Drug Shortages: A Complex Health Care Crisis." *Mayo Clinic Proceedings* 89 (3): 361–73.

Frank, Richard G., Andrew Hicks, and Ernst R. Berndt. 2021. "The Price to Consumers of Generic Pharmaceuticals: Beyond the Headlines." *Medical Care Research and Review* 78 (5): 585–90.

Frank, Richard G., Thomas G. McGuire, and Ian Nason. 2021. "The Evolution of Supply and Demand in Markets for Generic Drugs." *Milbank Quarterly* 99 (3): 828–52.

Galdin, Anaïs. 2023. "Resilience of Global Supply Chains and Generic Drug Shortages." Unpublished.

Ganapati, Sharat, and Rebecca McKibbin. 2023. "Markups and Fixed Costs in Generic and Off-Patent Pharmaceutical Markets." *Review of Economics and Statistics* 105 (6): 1606–14.

GAO (US Government Accountability Office). 2015. *Drug Shortages: Better Management of the Quota Process for Controlled Substances Needed; Coordination between DEA and FDA Should Be Improved.* GAO-15-202.

GAO. 2022. *FDA Should Take Additional Steps to Improve Its Foreign Inspection Program.* GAO-22-103611.

Grabowski, Henry G., and John M. Vernon. 1992. "Brand Loyalty, Entry, and Price Competition in Pharmaceuticals after the 1984 Drug Act." *Journal of Law and Economics* 35 (2): 331–50.

Hantel, Andrew, Mark Siegler, Fay Hlubocky, Kevin Colgan, and Christopher K. Daugherty. 2019. "Prevalence and Severity of Rationing during Drug Shortages: A National Survey of Health System Pharmacists." *JAMA Internal Medicine* 179 (5): 710–11.

He, Sijia, Sean Esteban McCabe, Rena M. Conti, Anna Volerman, and Kao-Ping Chua. Forthcoming. "Prescription Stimulant Dispensing to US Children: 2017–2023." *Pediatrics.*

HHS (US Department of Health and Human Services). 2024. "Fact Sheet: HHS Continues Taking Action to Increase Access and Supply of IV Fluids Following Hurricane Helene." https://public3.pagefreezer.com/browse/HHS.gov/02-01-2025T05:49/https:/www.hhs.gov/about/news/2024/10/18/fact-sheet-hhs-continues-action-increase-access-supply-iv-fluids-hurricane-helene.html.

HHS. 2025. *2005–2028 Draft Action Plan for Addressing Shortages of Medical Products and Critical Foods and Strengthening the Resilience of Medical Product and Critical Food Supply Chains.* US HHS.

Hopp, Wallace J., Lisa Brown, and Carolyn Shore. 2022. *Building Resilience into the Nation's Medical Product Supply Chains.* National Academies Press.

Howard, David H., Peter B. Bach, Ernst R. Berndt, and Rena M. Conti. 2015. "Pricing in the Market for Anticancer Drugs." *Journal of Economic Perspectives* 29 (1): 139–62.

HRSA (US Health Resources and Services Administration). 2013. "Statutory Prohibition on Group Purchasing Organization Participation—340B Drug Pricing Program. Release No. 2013-1." https://www.hrsa.gov/sites/default/files/hrsa/opa/prohibition-gpo-participation-02-07-13.pdf

HRSA. 2024. "The 340B Pricing Program." https://www.hrsa.gov/opa.

Huff, Charlotte. 2023. "Cancer Drug Shortages Deliver 'Gut Punch' to Patients Unsure If Their Survival Odds Will Be Undercut." *STAT News*, July 19. https://www.statnews.com/2023/07/19/cancer-drug-shortages-patients/.

IQVIA. 2023. *Drug Shortages in the U.S. 2023.* IQVIA Institute.

IQVIA. 2024. *The Use of Medicines in the U.S. 2024: Usage and Spending Trends and Outlook to 2028.* IQVIA Institute.

Jarvis, Lisa M. 2018. "Hurricane Maria's Lessons for the Drug Industry." *Chemical and Engineering News*, September 17. https://cen.acs.org/pharmaceuticals/biologics/Hurricane-Marias-lessons-drug-industry/96/i37.

Kaakeh, Rola, Burgunda V. Sweet, Cynthia Reilly, Colleen Bush, Sherry DeLoach, Barb Higgins, Angela M. Clark, and James Stevenson. 2011. "Impact of Drug Shortages on U.S. Health Systems." *American Journal of Health-System Pharmacy* 68 (19): 1811–19.

Kansteiner, Fraser. 2023. "Burn after Reading: FDA Blasts Intas for 'Cascade of Failure' after Investigators Find Heaps of Shredded Documents." FiercePharma, January 20. https://www.fiercepharma.com/manufacturing/burn-after-reading-fda-blasts-intas-cascade-failures-after-investigators-find-heaps.

Kansteiner, Fraser. 2024. "Baxter Offers Recovery Timeline after Major IV Fluid Plant Closure Due to Hurricane Helene." FiercePharma, October 9. https://www.fiercepharma.com/pharma/baxter-offers-recovery-timeline-after-major-iv-fluid-plant-closure-due-hurricane-helene.

Kim, Sang-Hyun, and Fiona Scott Morton. 2015. "A Model of Generic Drug Shortages: Supply Disruptions, Demand Substitution, and Price Control." Unpublished.

Manning, Richard, and Fred Selck. 2017. *Penalizing Generic Drugs with the CPI Rebate Will Reduce Competition and Increase the Likelihood of Drug Shortages*. Bates White Economic Consulting.

McBride, Ali, Lisa M. Holle, Colleen Westendorf, Margaret Sidebottom, Niesha Griffith, Raymond J. Muller, and James M. Hoffman. 2013. "National Survey on the Effect of Oncology Drug Shortages on Cancer Care." *American Journal of Health-System Pharmacy* 70 (7): 609–17.

McKesson. n.d. "McKesson's Rapid Returns Solution." https://www.mckesson.com/pharmaceutical-distribution/unsaleable-returns/ (accessed April 1, 2025).

Mulcahy, Andrew W., Preethi Rao, Vishnupriya Kareddy, Denis Agniel, Jonathan S. Levin, and Daniel Schwam. 2021. *Assessing Relationships between Drug Shortages in the United States and Other Countries*. Rand Corporation.

Nikpay, Sayeh, Melinda Buntin, and Rena M. Conti. 2018. "Diversity of Participants in the 340B Drug Pricing Program for US Hospitals." *JAMA Intern Medicine* 178 (8): 1124–27.

Noguchi, Yuki. 2023. "The Hospital Ran Out of Her Child's Cancer Drug. Now She's Fighting to End Shortages." NPR All Things Considered, October 23. https://www.npr.org/sections/health-shots/2023/10/23/1204856094/hospital-ran-out-child-cancer-drug-shortage.

Park, Minje, Anita L. Carson, Erin R. Fox, and Rena M. Conti. 2023. "Stockpiling at the Onset of the COVID-19 Pandemic: An Empirical Analysis of National Prescription Drug Sales and Prices." *Management Science* 70 (1): 6483–6501.

Park, Minje, Anita Carson, and Rena M Conti. 2023. "Linking Medication Errors to Supply Chain Disruptions: Evidence from Heparin Shortages Caused by Hurricane Maria." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4472407.

QuicksortRx. 2023. "Top Data to Track for Hospital Pharmacy Procurement Savings." August 8. https://blog.quicksortrx.com/top-data-to-track-for-hospital-pharmacy-procurement-savings.

Rees, Victoria. 2020a. "India to Restrict 10 Percent of Medicine Exports Due to Coronavirus." *European Pharmaceutical Review*, March 4. https://www.europeanpharmaceuticalreview.com/news/114484/india-to-restrict-10-percent-of-medicine-exports-due-to-coronavirus/.

Rees, Victoria. 2020b. "UK Bans Parallel Exporting of Crucial Medicines to Help COVID-19 Patients." *European Pharmaceutical Review*, March 23. https://www.europeanpharmaceuticalreview.com/news/115637/uk-bans-parallel-exporting-of-crucial-medicines-to-help-covid-19-patients/.

Schulman, Kevin A., Wasan M. Kumar, and Eli Y. Adashi. 2023. "Ensuring Access to Generic Medications in the US." *Health Affairs Forefront*, September 12. https://www.healthaffairs.org/content/forefront/ensuring-access-generic-medications-us.

Scott Morton, Fiona M. 1999. "Entry Decisions in the Generic Pharmaceutical Industry." *RAND Journal of Economics* 30 (3): 421–40.

Scott Morton, Fiona, and Margaret Kyle. 2011. "Markets for Pharmaceutical Products." In *Handbook of Health Economics*, Vol. 2, edited by Mark V. Pauly, Thomas G. Mcguire, and Pedro P. Barros, 763–823. North Holland.

Sivashanker, Karthik, John Fanikos, and Allen Kachalia. 2018. "Addressing the Lack of Competition in Generic Drugs to Improve Healthcare Quality and Safety." *Journal of General Internal Medicine* 33 (11): 2005–07.

Stomberg, Christopher. 2016. "Drug Shortages, Pricing, and Regulatory Activity." NBER Working Paper 22912.

Tucker, Emily L., Yizhou Cao, Erin R. Fox, and Burgunda V. Sweet. 2020. "The Drug Shortage Era:

A Scoping Review of the Literature 2001–2019." *Clinical Pharmacology and Therapeutics* 108 (6): 1150–55.

Tucker, Emily L., and Mark S. Daskin. 2022. "Pharmaceutical Supply Chain Reliability and Effects on Drug Shortages." *Computers and Industrial Engineering* 169: 108258.

US Centers for Medicare and Medicaid Services. 2024. "CMS Proposes New Policies to Support Underserved Communities, Ease Drug Shortages, and Promote Patient Safety." CMS Newsroom, April 10. https://www.cms.gov/newsroom/press-releases/cms-proposes-new-policies-support-underserved-communities-ease-drug-shortages-and-promote-patient.

US Department of Health and Human Services Office of the Secretary. 2024. *Policy Considerations to Prevent Drug Shortages and Mitigate Supply Chain Vulnerabilities in the United States.* US HHS.

US Senate Committee on Finance. 2024. "Wyden and Crapo Release Draft Legislation to Combat Prescription Drug Shortages." Newsroom, May 3. https://www.finance.senate.gov/chairmans-news/wyden-and-crapo-release-draft-legislation-to-combat-prescription-drug-shortages.

US Senate Committee on Homeland Security and Governmental Affairs. 2023. *Short Supply: The Health and National Security Risks of Drug Shortages.* US Senate.

USP (US Pharmacopeial Convention). 2024. "US Drug Shortages Reach Decade-High and Last Longer." USP, June 4. https://www.usp.org/news/us-drug-shortages-reach-decade-high-and-last-longer.

Vail, Emily, Hayley B. Gershengorn, May Hua, Allan J. Walkey, Gordon Rubenfeld, and Hannah Wunsch. 2017. "Association between US Norepinephrine Shortage and Mortality among Patients with Septic Shock." *JAMA* 317 (14): 1433–42.

Vizient. 2019. "Drug Shortages and Labor Costs." https://www.vizientinc.com/newsroom/news-releases/2019/new-vizient-survey-finds-drug-shortages-cost-hospitals-just-under-360m-annually-in-labor-expenses.

Wang, Yixin (Iris), Jun Li, and Ravi Anupindi. 2023. "Manufacturing and Regulatory Barriers to Generic Drug Competition: A Structural Model Approach." *Management Science* 69 (3): 1449–67.

Woodcock, Janet, and Marta Wosińska. 2013. "Economic and Technological Drivers of Generic Sterile Injectable Drug Shortages." *Clinical Pharmacology and Therapeutics* 93 (2): 170–76.

Wosińska, Marta. 2024. "Drug Shortages: A Guide to Policy Solutions." Brookings Institution, March 13. https://www.brookings.edu/articles/drug-shortages-a-guide-to-policy-solutions.

Wosińska, Marta. 2025. "Will pharmaceutical tariffs achieve their goals?" Brookings Institution. https://www.brookings.edu/articles/pharmaceutical-tariffs-how-they-play-out/.

Wosińska, Marta, Kalah Auchincloss, and Ilisa Bernstein. 2024. "FDA Oversight of Drug Manufacturing and Compounding: A Comparison." Brookings Institution, December 19. https://www.brookings.edu/articles/fda-oversight-of-drug-manufacturing-and-compounding-a-comparison/.

Wosińska, Marta, Rena M. Conti, and Elisabeth Reynolds. 2024. "Workshop Summary: Technology Solutions for Improving the Resilience of Generic Prescription Drug Manufacturing." Brookings Institution, January 11. https://www.brookings.edu/articles/workshop-summary-technology-solutions-for-improving-the-resilience-of-generic-prescription-drug-manufacturing.

Wosińska Marta E., Erin R. Fox, and Valerie Jensen. 2015. "Are Shortages Going Down or Not? Interpreting Data from the FDA and the University of Utah Drug Information Service." *Health Affairs Forefront*, April 8. https://www.healthaffairs.org/content/forefront/shortages-going-down-not-interpreting-data-fda-and-university-utah-drug-information.

Wosińska, Marta, and Richard G. Frank. 2022. "To Prevent Public Health Crises, We Need to Update the Essential Medical Product List." Brookings Institution, June 24. https://www.brookings.edu/articles/to-prevent-public-health-crises-we-need-to-update-the-essential-medical-product-list/.

Wosińska, Marta, and Richard G. Frank. 2023. "Federal Policies to Address Persistent Generic Drug Shortages." Brookings Institution, June 21. https://www.brookings.edu/articles/federal-policies-to-address-persistent-generic-drug-shortages.

Wosińska, Marta, and Richard G. Frank. 2024. "Comments on Senate Finance Committee Draft Legislation to Combat Prescription Drug Shortages." Brookings Institution, July 5. https://www.brookings.edu/articles/comments-on-senate-finance-committee-draft-legislation-to-combat-prescription-drug-shortages/.

Wosińska, Marta E., T. Joseph Mattingly II, and Rena M. Conti. 2023. "A Framework for Prioritizing Pharmaceutical Supply Chain Interventions." *Health Affairs Forefront*, September 13. https://www.healthaffairs.org/content/forefront/framework-prioritizing-pharmaceutical-supply-chain-interventions.

Yurukoglu, Ali, Eli Liebman, and David B. Ridley. 2017. "The Role of Government Reimbursement in Drug Shortages." *American Economic Journal: Economic Policy* 9 (2): 348–82.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 103–126

Measuring Income and Income Inequality

Conor Clarke and Wojciech Kopczuk

The distribution of income is an alluring topic for researchers, policymakers, journalists, and the public. The economics literature on measuring income inequality extends back at least to Kuznets (1953). But the contrast between two recent and prominent contributions by Piketty, Saez, and Zucman (2018) and Auten and Splinter (2024a) raises basic methodological issues that go to the heart of this literature. The two papers pursue the same goal—the measurement of United States top income inequality—but arrive at strikingly different conclusions: Piketty, Saez, and Zucman show a large increase in recent inequality, while Auten and Splinter show a pattern that is much more flat. In this paper, we use the controversy to return to first principles of measuring income and income inequality. We begin with a basic discussion of the income concept—starting with so-called comprehensive Haig-Simons income—and what it can and cannot capture. We then critically evaluate the income metrics in recent literature, the decisions that go into constructing those metrics, and the problems with underlying data sources.

Our goal is not to adjudicate the recent debate. Instead, we emphasize two themes about income and inequality measurement. First, no current measures of the income distribution use (or could use) the comprehensive Haig-Simons measure of income—the leading income concept that attempts to capture all changes in savings and consumption. Indeed, no measure could attempt to be truly comprehensive without making controversial methodological choices, and it is far from

■ *Conor Clarke is Associate Professor of Law, Washington University in St. Louis School of Law, St. Louis, Missouri. Wojciech Kopczuk is Professor of Economics and of International and Public Affairs, Columbia University, New York City, New York. Their email addresses are conor.clarke@wustl.edu and wojciech.kopczuk@columbia.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241424.

obvious that a comprehensive income measure would capture the most important distributional details that society should rightly value. The Haig-Simons measure of income is a common and influential ideal—and can be defended as a kind of necessary evil—but its virtues should not be overstated and its vices should not be forgotten.

Second, the efforts to quantify the elusive income concept raise tradeoffs. Papers that stick closely to what is observable in administrative tax data to measure income have many virtues, such as plentiful individual-level data and data that speak to the very top of the income distribution.[1] But the income that appears on annual tax returns reflects only a shrinking subset of a more comprehensive Haig-Simons concept. On the other hand, papers that attempt to use a more comprehensive (if still incomplete) income definition derived from the national accounts must extend beyond what can be observed on tax returns—and, often, any administrative data sources. Allocating these broader income concepts without individual administrative data requires contestable assumptions—which, in turn, helps to explain the recent controversy over the different findings of Piketty, Saez, and Zucman (2018) and Auten and Splinter (2024a).

In short, the ideal income concept is flawed, and attempts to approach this flawed ideal must come at a steep methodological cost. We conclude by sketching a few additional implications for the contemporary debate over income inequality and future literature.

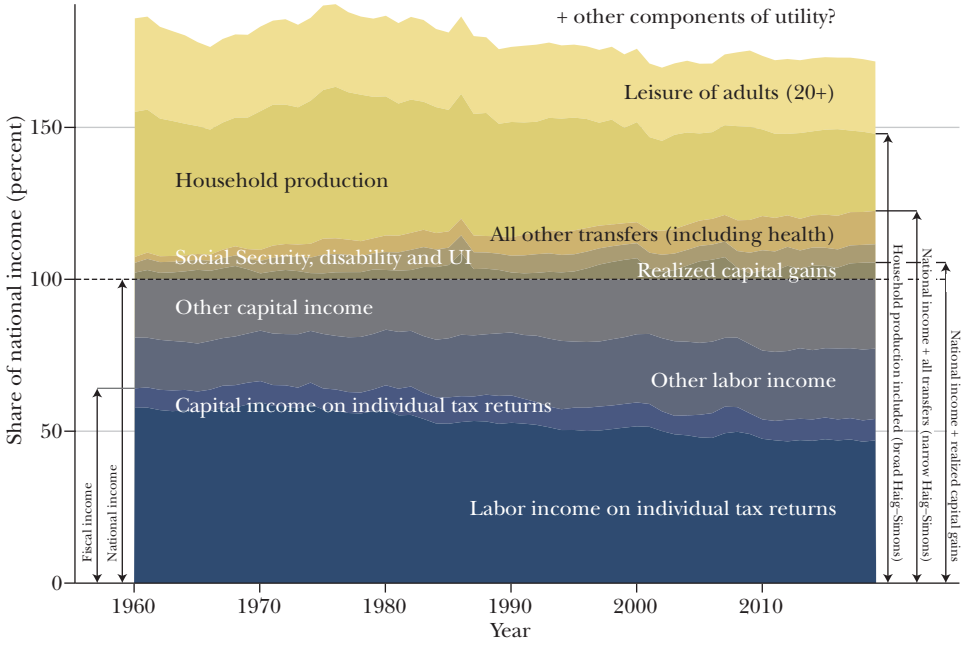
What Measure of Income Do We Want, and Why?

There is a bewildering profusion of income concepts: national, personal, labor, capital, factor, fiscal, cash, market, expanded, taxable, gross, net, pre-tax, post-tax, post-transfer, disposable—and many more. The sheer profusion of income concepts can give the impression that "income" results are driven as much by definitional choices as real economic changes. It is also not clear that *any* of these concepts align well with what nonacademics intuitively consider income. For example, Dahl and Ransom (2002) surveyed Mormons about their understanding of income for religious tithing purposes and concluded that it did "not coincide with current tax laws or economists' views of comprehensive income."

Figure 1 provides a sense of the relationship between various income measures and concepts. The bottom two categories show labor and capital income that appears on individual tax returns as a share of national income, as defined by US national accounts. Such income is a declining share of national income over time. The next level shows untaxed labor income—things like employer-provided health insurance and other fringe benefits—while the following level shows capital income that is not subject to individual taxation (and, hence, not directly linked to particular

[1] Because we focus on attempts to measure top income shares, tax data rather than survey data are our default for discussion.

Figure 1
Components of Different Income Concepts (National Aggregates)



Source: Piketty, Saez, and Zucman (2018), supplemented with data from the Internal Revenue Service, Congressional Budget Office, and Bureau of Economic Analysis. See online Appendix for details.
Note: The figure shows a breakdown of components of different income concepts, expressed as shares of national income.

individuals). This category of capital income includes a variety of gains that the US tax system does not target, such as the rental value of owner-occupied housing and the interest from certain tax-advantaged retirement accounts. It also includes corporate retained earnings and corporate tax liability itself. These forms of income are included in national income as measured by our national accounting system—every component of which is included under the horizontal line at 100 percent.

But national income is not everything. The categories above the national income line offer a sense of forms of individual economic gain that US national accounts do not include. The first category shows realized capital gains, which are a familiar part of US tax law—you pay tax on an appreciated asset when you sell it—but are not part of "national income" as officially defined.[2] The next category shows another familiar form of economic benefits: payments from Social Security,

[2] The precision of this figure should be not be overstated: some components of asset appreciation—such as corporate retained earnings—are included in the national accounts, so that adding realized capital gains may double-count some gains. In addition, the measure on the graph does not directly include tax-exempt housing capital gains and it does not account for inflation.

disability insurance, and unemployment insurance. The category above that shows all other government transfers, including those for healthcare spending. The sum of those categories, plus national income, can be viewed as a "narrow" version of the comprehensive Haig-Simons definition. The next category above shows the value of unpaid household production—which is based on surveys of time-use and included in official "satellite" national accounts. Adding household production to the categories below it creates a total that can be viewed as closer to a "broad" Haig-Simons definition. But adding other categories is possible, too. The top category includes a rough and illustrative value of leisure (based on time-use survey averages and multiplied by an indexed version of the federal minimum wage).[3] One could also imagine adding yet more hard-to-value components of welfare on top of that (sleep, happiness, health, and the like). This list of categories is intended to be provocative more than precise. But we hope that it illustrates both the fuzzy contours of the income concept and the large differences between aggregate definitions.

The Haig-Simons Baseline

Where does our modern, baseline concept of "income" come from? The so-called "Haig-Simons" comprehensive definition of income dates to the writings of economists Robert Haig (1921) and Henry Simons (1938). Haig defined income as "the money value of the net accretion to one's economic power between two points of time," and Simons defined it as "the algebraic sum of (1) the market value of rights exercised in consumption and (2) the change in the value of the store of property rights between the beginning and end of the period in question." Broadly and somewhat informally, the Haig-Simons concept—sometimes shorthanded as "economic" income—defines the income of an individual as the market value of all changes in the ability of that individual to save or consume over a given period.

Haig and Simons were writing in the wake of the ratification of the Sixteenth Amendment (1913), which empowered the US Congress "to lay and collect taxes on incomes, from whatever source derived."[4] The backdrop for their pioneering work was thus an early administrative income tax system for which the basic concepts—like the appropriate scope of "taxes on incomes"—were still being worked out. The concept of "income" existed before the invention of an income tax, of course. But, prior to its use as a tax base, "personal income" was (to our knowledge) not a central concept in scholarly work. In Adam Smith's *The Wealth of Nations* (1776), for example,

[3] The online Appendix describes the construction of this figure in more detail.

[4] More specifically, the Sixteenth Amendment allowed Congress to tax incomes without apportioning such a tax on the basis of state population—that is, such that a state with 5 percent of the national population would bear 5 percent of the total tax burden. In 1895, in the landmark case of *Pollock v. Farmers' Loan and Trust Company* (157 U.S. 429 [1895]), the US Supreme Court held that an income tax needed to be apportioned by population. Because an apportioned income tax cannot guarantee the equal treatment of individuals with the same income in different states—or progressive rates—*Pollock* was a fundamental obstacle to US income taxation that led to the eventual ratification of the Sixteenth Amendment.

the concept of "income" makes only a vanishingly small number of appearances. Smith's discussion of taxation acknowledges that individuals should contribute "in proportion to the revenue which they respectively enjoy under the protection of the state." But he considered such taxes impossible. For Smith, "[t]he impossibility of taxing the people, in proportion to their revenue" is what led to "the invention of taxes upon consumable commodities"—which were more easily measurable and, Smith argued, will be "nearly in proportion to [one's] revenue." Early and temporary US experimentation with an income tax during the Civil War encountered a similar issue: "The government had no scientific way to measure personal income" (Brownlee 2016).

Haig and Simons were thus concerned with fashioning a new concept of income that would be relevant and administrable for both lawyers and policymakers. They had only a limited practical and philosophical inheritance upon which to draw—though their discussions did mirror some earlier German-language discussions of the income concept that flowed from European experimentation with income taxation (for example, von Kleinwächter 1896).

Haig (1921), for example—leaning on a tradition of early utilitarian thinkers—emphasized that an abstract and ideal economic analysis might recognize income as "a flow of satisfactions, of intangible psychological experiences." But he quickly conceded that practical analysis required "something more definite and more homogeneous—less diaphanous and elusive than these psychic satisfactions." For this reason, Haig argued that a usable concept of income must look to gains that could be readily converted into market prices. "The basis of comparison," he argued, "is, of course, that of the common, universally-acceptable unit of value—money." Simons (1938), likewise, argued that a workable definition of personal income "has to do not with sensations, services, or goods but rather with rights which command prices (or to which prices may be imputed)."

Over time, the Haig-Simons concept has obtained implicit dominance in tax law and tax policy. Many introductory legal and economics textbooks introduce students to the concept of Haig-Simons income as a baseline for thinking about the difference between "economic" and "legal" income, and for assessing individual provisions of tax law (for example, Graetz and Alstott 2022; Gruber 2022, p. 538). National governments rely on the Haig-Simons concept too. Each year, for example, the federal government estimates the size and costs of various "tax expenditures" (the name for quasi-spending programs that run through the tax code) and clarifies that such estimates use the baseline of a "comprehensive income tax, which defines income as the sum of consumption and the change in net wealth in a given period of time"—that is, Haig-Simons income. The Haig-Simons framework also underlies the guidelines of the United Nations Statistics Division, which provides standards for how statistical agencies around the world define income (Larrimore et al. 2021).

To be sure, there are excellent reasons for an income tax system to express some aspirational interest in capturing all changes in an individual's ability to save or consume. A more naïve concept of income—say, "cash received"—would quickly fall victim to an even more sophisticated world of fringe benefits: tax-free meals,

housing, insurance, and so forth. One defense of the broad Haig-Simons concept is therefore that it expresses the ambition to capture all forms of real economic gain—such as benefits paid in-kind—rather than mere salaries paid in cash or cash gains realized upon the sale of an asset.

But there are also serious drawbacks to the Haig-Simons concept, many of which have been known since the beginning and have been emphasized in recent research (as in Brooks 2018; Smeeding, Johnson, and Citro 2024). In the rest of this section, we discuss five interrelated problems that have particular relevance for the measurement of income and income inequality. First, Haig-Simons income is perhaps best conceptualized as an aspirational concept for administrative tax law, not a concept developed with justice or fairness in mind—and its proper role in debates over justice is debatable. Second, there are serious conceptual difficulties about what the comprehensive Haig-Simons concept can and should include, which raises stubborn questions about its implementability. Third, even where there is conceptual consensus about what to include, many forms of savings and consumption are inherently difficult to value. Fourth, the Haig-Simons concept does not tell us what the appropriate unit of observation is (individuals? households?) for measuring income and inequality. Finally, the Haig-Simons concept does not tell us how to attribute government spending to personal income (especially given that government spending was already income for someone else before it was taxed and spent).

Haig-Simons and Fairness

Haig and Simons were not writing about personal income in the context of fairness and equality; they were writing about income in the context of a nascent tax system attempting to find its footing. This meant the income concept that Haig and Simons developed already started with concessions to reality—such as Haig's comment that a more ideal utilitarian concept would concern itself with the "flow of satisfactions" and "intangible psychological experiences," rather than gains that could be readily translated into market prices. Other measures beyond "utility"—like some measure of welfare, flourishing, happiness, capability, opportunity, ability to pay, and so forth—would have a similarly imperfect overlap with the Haig-Simons concept.

One such concession to reality is that the measurement of Haig-Simons income limits itself to a single administrative time period (in the case of the US tax system, normally one year). Again, there is a good reason for this: Much of collective human life is conducted on an annual basis. But justice and fairness should care about more than a year. Income changes over the life cycle (for some discussion of inequality in this context, see Auerbach, Kotlikoff, and Koehler 2023). It is not necessarily a social fairness problem if the 20 year-old college student has a low Haig-Simons income in a given year. Income can also be lumpy. It is not necessarily a distributional problem if that college student grows up and, in one year, earns a one-time bonus or inherits a house.[5]

[5] Similar issues arise concerning whether the nation-state is the best level of generality in which to study income. Again, much collective life is conducted through nations. But it might be that local or global

Of course, the concept of Haig-Simons income does have *some* relationship to justice and fairness. Ideally, the tax system cares about measuring income because it cares about measuring ability to pay, and this particular concept of income is constructed with a defensible notion of ability to pay in mind. As anticipated centuries ago by Adam Smith, the tax system should care about measuring ability to pay because it reflects a fundamentally fair and reasonable way to allocate the cost of government. But "income"—as defined by Haig and Simons—is at best a rough indicator of the other measures of justice listed above, and it is an open question whether first principles of justice should care about one's time-restricted saving and consumption divorced from its impact on those other measures. Accordingly, the standard concept of "justice" used in economics (and based on a utilitarian welfare function) does not intersect directly with Haig-Simons income.

Perhaps the best defense of using Haig-Simons as a baseline for thinking about distributive issues goes like this. First, it makes sense to use income data to study distributional issues because we *have* income data—both from the work of the administrative tax system and from our system of national accounts. Second, while income is an imperfect way of measuring fairness, we have some reason to believe that it is connected to the measures we care about, because it relates to the ability to pay (as discussed above and by, for example, Musgrave 1967; Thuronyi 1990). Finally, every measure of social welfare or distribution has its faults (for example, Deaton 2020; Glogower 2023). Still, the particular faults of income as a rubric for thinking about justice and fairness—it was only partially intended to capture those things, and it only captures them imperfectly—should not be forgotten.

Haig-Simons and Implementability

How feasible is it to implement a broad and comprehensive concept of income? A fundamental problem is that that the concepts of "consumption" and "savings" (or changes in wealth) are not obvious or self-defining. Indeed, even in cases where there *is* rough consensus about the concepts, there can be uncertainty in how the concepts can be measured in common units of exchange that a tax system (and the Haig-Simons concept) require. The issues here are vast—indeed, defining and policing the boundaries of those concepts is a primary object of tax law and policy—and we provide here only a selective overview of some of the more fundamental issues.

One basic question is how to think about the frontier between personal consumption and expenses that represent costs of producing income. It is crucial to an income tax base—as opposed to a tax on the volume of activity or receipts—that it nets out the costs of producing income. If two business owners each bring in $100,000 of gross receipts in a given year, but one has $90,000 in expenses and the other has $10,000 in expenses, an income tax should not treat those owners as equivalent. If it did, the tax would be one on gross receipts, not income. Indeed,

inequality is what matters more. For some discussion of this issue, see Hayashi (2023).

many taxes (including many US taxes) were, and sometimes still are, imposed on gross receipts rather than on net income—presumably because gross receipts are easier to measure, even if they do less to capture ability to pay.

But figuring out what is a cost of business—as opposed to an item of personal consumption—turns out to be surprisingly tricky. (For some evidence of businesses blurring the line between costs and personal consumption, see Alstadsæter, Kopczuk, and Telle 2014; Leite 2024.) As Simons observed almost 100 years ago, a good or service can obviously be either, depending on the context. A professional artist who buys paint can treat it as a business expense, while a hobbyist who buys the same paint treats it as consumption. For Simons, this dual quality suggested an unsettling conclusion: "There is something quite arbitrary about the distinction between consumption and accumulation."

This issue is not a minor one. Just as the same activity may be viewed reasonably as representing consumption or production—depending on the context—so too may the same good or activity be shot through with both consumption and production. The activities most fundamental to the human experience—food, sleep, shelter—are both an obviously major component of consumption and a necessary condition of earning income. Indeed, the US tax system is somewhat schizophrenic in whether it treats such expenses as consumption or as a cost of production. For example, the deductibility of certain personal medical expenses may be viewed as either a departure from the normal income tax baseline (and thus a form of consumption we decline to tax) or as a reasonable part of the normal income tax baseline, because it does not tax a cost of labor production (that is, we need to be alive to produce income).

Or consider the issue from the other direction: Many jobs are enjoyable, and in that sense can be (partially) conceptualized as substantial consumption opportunities. A young economics or law student may decide to pursue the academic life—under the impression that it will be fun—over a career in finance or at a law firm, even if the latter may yield a far higher salary. Again, the issues here are conceptual, practical, and moral. How should we think about taxation of those who choose the "fun job," although their talents could have earned higher market returns in different professions? Should the tax system create incentives for individuals to pursue a "calling" that pays a lower income, or should it attempt to reach talents and abilities directly? (For some discussion of these issues in both economics and in law, useful starting points include Lockwood, Nathanson, and Weyl 2017; Zelenak 2006.)

Imputations and the Problem of Valuation

Both Haig and Simons imagined a concept of income that would add up "the money value" (in Haig's terms) or "the market value" (in Simons's) of changes in consumption and saving. But even when we can agree on the right conceptual bucket for an activity (see above), many large components of consumption and savings do not have readily ascertainable market values. Again, this raises a mix of issues that are both practical and philosophical. For example, the change in value of corporate

equities constitutes a large fraction of annual Haig-Simons income, but there is (by definition) no public market for closely held corporations. Or consider: Leisure can be thought of as a large component of consumption—and indeed working hours and vacation time are often in employment contracts—but there is no agreed-upon method for valuing moments of idle reverie in which one does nothing (although Figure 1 does provide one rough attempt).

The so-called "problem of valuation" is closely related to a recurring issue in the income and income inequality literature over what "imputations" of income can and should be done (where "imputation" in this context refers to an assignment of value that must be done indirectly). Most observers would agree, for example, that in thinking about the distribution of Haig-Simons income we should impute rental income to owners who occupy their own houses, on the theory that occupying a home increases one's ability to save or consume—just as having an employer cover your rent, or owning a home and renting it out as business, can increase one's ability to save or consume. But practice diverges: US tax law does not impute (and thereby attempt to tax) rental income to owner-occupiers, while US national accounts do attempt to include this as part of "national income."

What imputations of value are possible and where should one stop? Perhaps, for example, we could and should impute the income of businesses as it is earned to the owners of those businesses—as our tax system does in the case of some (but not all) businesses, and as we discuss in Clarke and Kopczuk (2017) and further below. This discrepancy is broadly related to another major deviation between the Haig-Simons concept and the tax system in the United States: the timing of when capital gains income is recognized by the tax system, or deemed "realized" (which generally happens when an asset is sold, rather than when it appreciates in value). As a legal matter, the status of realization remains contested.[6] As a policy matter, one traditional argument in favor of a realization requirement is that some assets are too difficult to value on an accrual basis.

Then there is household production. Many tasks could in principle be outsourced to the labor market—cooking, cleaning, driving, childcare, and so on. Just as one might think about imputing rental income for one's own house, we might think about imputing labor income for one's own labor. This is not just a toy philosophical issue. Household labor consumes a lot of time—hundreds of billions of hours each year in the United States (Bridgman, Craig, and Kanal 2022)—and correlates with other elements of Haig-Simons income. Household labor can also change dramatically over time—as we have seen in both recent decades (as the composition of the workforce has changed) and in recent years (as the pandemic temporarily changed labor markets). There is no consensus on how to account for

[6] In the high-profile 2024 Supreme Court case of *Moore v. United States* (602 US 572 [2024]), four of the nine justices indicated their view that "realization" was a constitutional requirement—in other words, that income must be "realized" in some way before the tax system can reach it under the authority of the Sixteenth Amendment. For present purposes, the opinion illustrates that there may be important and ongoing differences in how economists and the legal system think about "income."

household production—and our tax system and standard national accounts do not attempt to account for it. However, satellite estimates of gross domestic product suggest that including household production would add more than 25 percent to standard measures of GDP in 2020 (Bridgman, Craig, and Kanal 2022).[7] The income inequality literature is only starting to address this issue, and Figure 1 provides some sense of its potential scale. Incorporating household production into inequality measures could have a large effect on measured levels and trends (Frazis and Stewart 2011; Gautham and Folbre 2024).

The Proper Unit of Observation

In their initial discussions, Haig and Simons were concerned with the "personal" income of individuals. But it is not obvious what the best unit of analysis is, and tax systems vary in how they approach this issue. For example, for the first few decades, the US income tax system only looked to individual income and did not, as now, allow varieties of joint (household) filing to occur. (Indeed, at certain junctures this created an incentive for a household to attempt to reduce its taxes by formally splitting a husband's income with a nonworking wife, on the theory that they would each pay a separate individual tax at a lower marginal rate.)

For normative and distributional analysis, there is certainly a case for looking to individuals. Much of life concerns individuals. We *are* individuals. It seems obvious and reasonable to consider the gains and capacities of individuals. But other facts of life seem obvious and reasonable, too. Individuals organize themselves into households, extended families, and even communities to pool resources and pursue shared plans. It therefore seems reasonable to think about both the "income" of individuals and of households—and indeed potentially of larger units, such as extended families, communities, and so forth. On the other hand, households form, change, and dissolve over a lifetime, which may push in favor of using individuals as the right unit of measurement. The naked concept of "income" generates no consensus on the *right* way to think about these issues—although any administrative tax system and any discussion of inequality must necessarily take a stance on them.

[7] Whether household work should be included in national production and national income has been an issue since the very beginning. Here is the rather conclusory and unsatisfying discussion of the issue in Department of Commerce (1934), which was the government's original report on the national income concept developed under the guidance of Simon Kuznets: "The volume of services rendered by housewives and other members of the household toward the satisfaction of wants must be imposing indeed, when totaled for the 30 million families comprising the population of this country; and the item is thus large enough to affect materially any estimate of national income. But the organization of these services render them an integral part of family life at large, rather than of the specifically business life of the nation. Such services are, therefore, quite removed from those which gainfully occupied groups undertake to perform in return for wages, salaries, or profits. It was considered best to omit this large group of services from national income, especially since no reliable basis is available for estimating their value. This omission, unavoidable though it is, lowers the value of national income measurements as indexes of the nation's productivity in conditions of recent years when the contraction of the market economy was accompanied by an expansion of activity within the family."

Government Spending and Income

From the Haig-Simons perspective, some government spending may seem to translate easily into income. Someone who receives a Social Security check, for example, has money to save or consume (and Social Security benefits generally contribute to taxable income above a certain level). But even this seemingly simple case disguises great complexity. After all, Social Security benefits correspond to taxes that somebody else pays. Including both taxes paid and benefits received in a measure of income would involve double-counting. We might be tempted simply to deduct Social Security taxes from income. But that would diverge from one administrative purpose of the underlying income concept—namely, measuring one's ability to pay for the costs of government.

How other government spending translates into ability-to-pay poses even harder challenges. Paved roads paid for by a government also increase one's ability to save or consume, but it is harder to attribute the resulting income to individuals or households. Similarly, national defense supports the presence of relatively secure property rights, and the absence of pillaging invaders improves savings and consumption. But it is difficult to say how such forms of spending should be allocated to households and individuals across the income spectrum, if at all. The size of such benefits—and how fast the benefits rise in proportion to other forms of income—have been debated by economists, philosophers, and lawyers for decades (for example, Blum and Kalven 1952; Murphy and Nagel 2002; Piketty, Saez, and Zucman 2018).

The question begins to approximate: What is government worth to each individual? But, as noted above, most tax systems are concerned with estimating one's ability to pay *for* the costs of government, not with estimating how the costs of government improve one's ability to save or consume. Moreover, the issues of how government taxing and spending are interrelated with income are not easily resolved by appealing to *both* a pre- and a post-tax perspective, as is the case in recent work allocating national income by both Piketty, Saez, and Zucman (2018) and Auten and Splinter (2024a). This recent work roughly allocates *current* government spending—already a formidable and contestable imputation task, as discussed more below. But it does not attempt to impute income—which would be akin to a kind of imputed rental income—from the myriad forms of capital (both physical and institutional) that the government owns, represents, or controls.

These five issues are neither trivial nor new—much less resolved. Many of them were reflected in the early German-language discussions of the income concept, were repeated by Haig and Simons, and were raised again in the debates in the 1960s and 1970s over the concept of "tax expenditures" and the development of a national tax expenditure budget (for example, Bittker 1967). Similar concerns have been raised in recent legal scholarship (as in Brooks 2018). But they may have been forgotten in some of the recent, public debates over the distribution of income—which have treated income as both an obvious and obviously meaningful unit of measurement and dimension of comparison.

Recent Controversies over US Income Inequality

In recent studies of US top income inequality, two broad families of "income" measure are in wide use: income measures that start with tax data and income measures that start with national-accounts data. However, the measures of income that reach beyond what can be directly observed in administrative tax data have to rely on extensive imputations based on other data sources. These other data sources can include aggregates, surveys, smaller-sample estimates, and other administrative information that cannot be directly linked to observed income. Sometimes, evidence may only be available for a select number of years, so researchers must project to other periods. Sometimes, there is not much information at all.

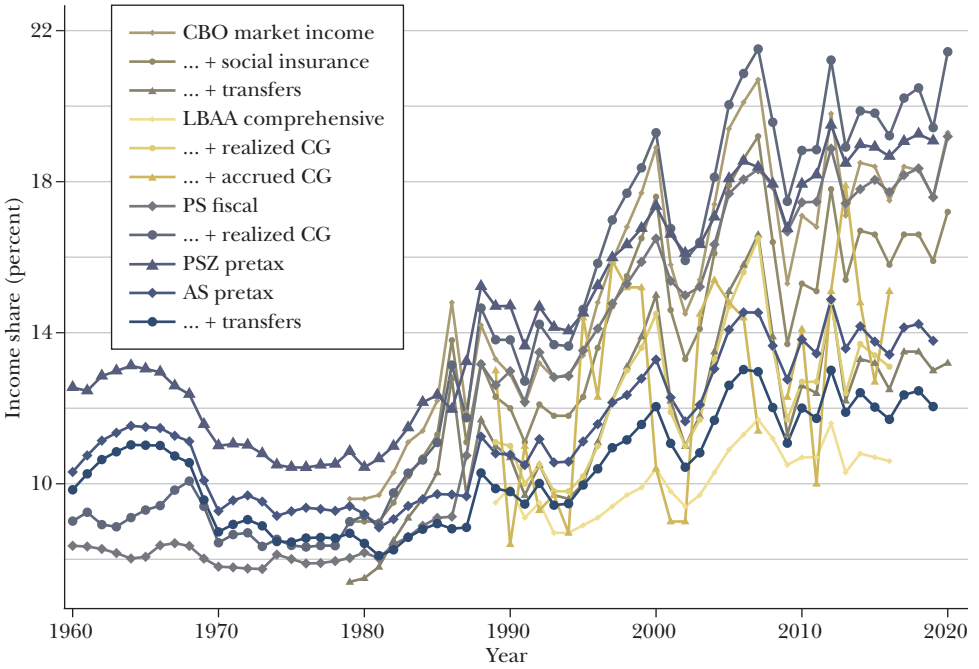
These data limitations require many assumptions that leave room for subjective judgment calls. It is therefore unsurprising that different researchers pursuing nominally the same exercise will reach different answers. At the same time, the different answers put consumers of such research in the unenviable position of trying to evaluate the underlying judgment calls—often with no consensus methodology on which to rely. Sometimes, the best a casual reader can do is rely on the authors' own summaries of the differences.

Figure 2 provides an overview of how different choices and definitions produce different distributional results. It displays the top 1 percent income shares over time drawn from some prominent sources in both government and the academic literature. Most of these series are pre-tax, although a number of the measures include some government transfers. The series from Larrimore et al. (2021) nets out taxes, but is included here because they are unique for their detailed accounting for capital gains. Given the ambiguity and flexibility of the income concept, the variations in the levels and trends are, unsurprisingly, large. In recent years, for example, the measured top 1 percent share of income varies between almost 25 percent and less than 9 percent, depending on which measure of income is selected. The figure may be a little messy and disorienting. But that is the point: the lack of consensus produces disorienting results.

There are several main sources here. Three estimates are from the Congressional Budget Office, based on household data: one for market income, one adding just social insurance transfers (like Social Security), and one adding all transfers. Several of the results are drawn from studies by Piketty and Saez, sometimes with Gabriel Zucman. For example, Piketty and Saez (2003) uses tax units as data in one line, and then add realized capital gains in a second line. Piketty, Saez, and Zucman (2018) uses individuals as the unit of observation and use national income as the basis for analysis. Another set of results is drawn from Larrimore et al. (2021), who start with a series that accounts for cash transfers and in-kind private and public benefits, then nets out taxes and adds both realized and accrued capital gains. Finally, we show two measures from Auten and Splinter (2024a): their measure based on pre-tax national income and a second that includes government transfers.

We do not think that any of these approaches is "right" or "wrong," and we do not think that there is a general-purpose series that one should pick above the rest.

Figure 2
Different Estimates of the Top 1 Percent Share in the Literature



Source: Estimates from Larrimore et al. (2021) (LBAA), Congressional Budget Office (2023) (CBO), Piketty and Saez (2003) (PS, updated series), Piketty, Saez, and Zucman (2018) (PSZ, updated series), and Auten and Splinter (2024a) (AS).
Note: The figure shows eleven different measures of the share of US income earned each year by the top 1 percent of income-earners.

The estimates correspond to different concepts of income and different definitions of the population, and they vary along other dimensions, too. For example, some authors vary in whether they include payroll taxes or Social Security benefits in the pre-tax concept of income, or in how they account for tax evasion, or in how they distribute business incomes that are not easily attributable to individuals or households using administrative tax data. Given the varying concepts of income, the varying notions of the income-earning unit, and the varying data limitations, there are important methodological choices to make.

Nevertheless, some general patterns are discernible. How one accounts for capital gains matters. They are a large and volatile component of income—even more so when one considers them as they accrue rather than as they are realized. Whether and how one includes social insurance and transfers matters quite a bit too, as one might expect. Finally, there appears to be less disagreement before the mid-1980s than after—or, in other words, the trends vary. One striking aspect of these trends is that the pre-1980s series are smoother and less cyclical, which suggests

to us that the earlier series may not be accounting fully for business incomes—something that we will come back to and which we explore in other work.

The Tax-Law Income Baseline

The influential efforts by Piketty and Saez (2003) to measure US income inequality relied on data generated by the individual income tax, a concept they described as "fiscal income." This approach had, and continues to have, several strong practical justifications. First, prior literature that relied on survey data had incomplete coverage at the very top of the income distribution. But, at least in principle, the modern tax system sees virtually all income-earners. Second, the tax data enable (with some assumptions) both the observation of individual units and the construction of national aggregates. Third, US income tax data go back a long time—in some cases, close to the ratification of the Sixteenth Amendment in 1913. In principle, therefore, tax data allow for the construction of long-run series on "income" that is rich and granular along many dimensions.[8]

But relying on tax data results in large deviations from the baseline Haig-Simons income concept. As Atkinson, Piketty, and Saez (2011) acknowledged, all the income estimates in this literature "follow the tax law, rather than a 'preferred' definition of income, such as the Haig-Simons comprehensive definition." As a result, researchers who use tax data to measure "income" inequality are not actually measuring what the "income" terminology often suggests. Moreover, we should expect the size and nature of the difference between a "preferred" income measure and one derived from tax law to evolve over time—as both tax law and the real economy evolve.

One glimpse of these points is visible back in Figure 1, which illustrates that taxable labor and capital income is an incomplete share of national income and other income concepts, and that this share changes over time. As Piketty, Saez, and Zucman (2018) note, "the fraction of national income that is reported in individual income tax data has declined from 70 percent in the late 1970s to about 60 percent today." Why is that? Here, we recount a few of the particularly large divergences between national income and tax-law income.

First, the United States taxes many forms of economic income only when they are "realized," rather than when they accrue. For example, a family may own real property, equities, and other important assets that increase in value over the course of a year—thereby increasing that family's ability to save or consume. But the US tax system does not see taxable income until (for example) an asset is exchanged or sold—a so-called "realization event." Realization is a tax-law term of art that has

[8] Not everyone has to file a tax return. But, nowadays, tax authorities also have some information about the income of nonfilers from "information tax returns," which are required from some to provide information about taxes that are owed by others (say, a firm that hires an independent contractor). This information allows for estimating the aggregate income of nonfilers and—when coupled with the assumption that nonfilers are at the bottom of the income distribution—still allows for estimating top income shares. But this issue becomes thornier the further back in time one goes, both because the extent of information reporting has changed over time, and because the income tax system was far from universal until after World War II.

no conceptual analog in finance or economics. It is not just equivalent to the sale or disposition of an asset, or receipt of cash. It is not always intuitive; for example, there are cases where a person can exchange one asset for another of a similar type without triggering income tax liability. The realization requirement represents an enormous deviation from an ideal economic concept of income and has been famously characterized as the "Achilles heel" of the US income tax (Andrews 1983).

Second, the United States separately taxes individual income (via an individual income tax) and certain corporate income (via a corporate income tax). Income earned by corporations that pay the corporate income tax is not taxed to individuals until it is distributed to them, or until individuals sell their equities and "realize" income. On the other hand, the income of many "pass-through" entities is immediately attributed to (and thus individually taxable to) their owners. The separately taxed corporate sector—and the shifting legal line between the individual and corporate sectors for tax purposes—raises many complex issues for distributional work that relies on tax data (for an overview, see Clarke and Kopczuk 2017).

Third, and related, personal income on individual tax returns does not account for business and employer-side taxes (like the corporate income tax and an employer's share of payroll taxes). This means that the income appearing on individual tax returns cannot be used to construct an airtight measure of "pre-tax" income. Individual pre-tax income is already net of business-side taxes.

Fourth, through the "tax expenditure" policies mentioned earlier, the United States excludes from taxable income many large items that uncontroversially increase the ability of individuals to save or consume (for an overview, see Joint Committee on Taxation 2022). For example, the value of employer-provided healthcare plans is generally excluded from taxation. As another example, the basis of property acquired at death—say, inherited stock—is stepped up to its "fair-market value," meaning that untaxed appreciation will escape taxation altogether. These tax expenditure policies are numerous and large.

Fifth, income that is subject to tax evasion and tax avoidance is not fully accounted for in a tax-based measure of income. The income of nonfilers is also not on tax returns, although Piketty and Saez (2003) do estimate and add this as a component of income in the aggregate.

Despite these issues, the use of tax data to study distributional questions has a strong rationale. After all, tax law reflects a particular society's answer (at a particular time) to the difficult and inevitable questions about ability-to-pay that have been known since the time of Haig and Simons. Moreover, especially at the top of the income distribution, this is the best individual data on income that we have.

But this justification must come with caveats. Any income measure derived from tax law will be missing important forms of economic gain—which will, in turn, vary systemically across the income distribution. For example, we should expect taxpayers to differ systematically across income levels in the degree to which they benefit from the realization doctrine, own corporations, and receive large employer-provided healthcare plans. Legal definitions also change over time. Large legal changes—like the major statutory tax reform that happened in 1986—cast doubt on the ability of

law-derived measures to produce a consistent measure of income. Finally, concepts of income derived from tax law will vary across countries—casting doubt on the power of international comparisons, as well as historical ones.

National Accounts and Opposing Views

The scale and nature of the deviations between tax-law-based income measures and a more comprehensive definition of income led to a shift in the research agenda on income inequality: a turn from nearly exclusive reliance on tax data to instead starting analysis with national-accounts data. Thus, Piketty, Saez, and Zucman (2018) attempted to address some of the drawbacks of the earlier Piketty and Saez (2003) approach by allocating national income as defined by the national income and product accounts from the US Bureau of Economic Analysis. But this approach also has drawbacks.

For one, the national-accounts definition of income—while broader than tax-derived measures of income—is not the same as comprehensive Haig-Simons income. This point is typically not emphasized (or even mentioned) in the economics literature, but it is important. Of course, national-accounts income does include many classic forms of economic income that our tax system does not attempt to reach—such as the imputed rent of owner-occupied housing (Mayerhauser and Reinsdorf 2007). But those imputations go only so far. For example, the national accounts do not attempt to impute household labor production, and they do not account for unrealized increases in asset values. Once again, a sense of the magnitude of this issue is provided by Figure 1.

There is also an important practical tradeoff between measures that start with tax data and national-accounts data. The attempt to approach—if never quite arrive at—a comprehensive definition of income comes at the expense of many of the practical virtues described above. It is intrinsically difficult to allocate income to individuals that is not reported on individuals' tax returns. Researchers must instead resort to imputations based on imperfect information and debatable assumptions, which inevitably leads to disagreements. Mirroring our discussion above, we offer an illustrative (and by no means exhaustive) list of six categories in the national accounts without reliable individual data.

A first example: undistributed corporate earnings are included in national income; realized capital gains are not. Undistributed corporate income is part of the value of corporate equities owned by individuals, but it is not synonymous with accrued capital gains and certainly not with realized ones. Without good data to link firms and individuals—and such data do not, to our knowledge, exist in the United States—researchers need some method for imputing undistributed corporate income to individuals. When starting with individual tax data, researchers can rely on proxies for ownership, like dividends and capital-gains realizations. But firms that pay dividends are not necessarily the same firms that retain earnings (after all, paying dividends *reduces* retained earnings). Indeed, the relationship between retained income and payouts is likely heterogeneous over time and across firms. Legal changes like the 1986 Tax Reform Act—which massively expanded the

number of "pass-through" corporations and moved whole categories of firms off the corporate income tax—surely did not have a neutral effect on the relationship between dividends and corporate retained earnings. Relying on dividends to impute corporate retained earnings does poorly in Norway, where more data are available to link firms and owners and thus study this connection (Alstadsæter et al. 2016). Similarly, evidence from Honduras suggests most income at the top of the distribution consists of undistributed corporate profits, even though most individuals at the top do not receive dividends (Del Carmen et al. 2023).[9]

Second, savings accumulate in pension funds and retirement accounts, but researchers do not, at present, observe the ownership of these accounts. What should be imputed to individuals, and when? Do retirement savings generate Haig-Simons income at the time of contribution and accrual, or the time of withdrawal after retirement? Most approaches treat retirement savings as income when they are distributed from retirement accounts. But this approach raises a practical problem: national accounts measure retirement savings on an accrual basis and—if the researcher is committed to allocating the full annual aggregates that appear in the national accounts—the difference between accrual and distribution must be allocated. Retirement income is also behind one of the mistakes that crept up when dealing with imperfect data—as Auten and Splinter (2024a) document, the approach of Piketty, Saez, and Zucman (2018) misinterpreted some rollovers of retirement funds to different IRA and 401(k) accounts as new income—because these mere transfers of assets from one account to another do appear on tax returns. Because large rollovers are concentrated at the top of the income distribution, this had the effect of increasing top income shares.

Third, reliance on national accounts requires choices about the treatment of Social Security—and, more generally, blurs the line between the nominally pre- and after-tax approaches to measuring income. Social Security transfers are funded out of payroll taxes, but they also constitute a component of income that is partially subject to taxation. As noted above, including both the taxes and benefits would create double-counting. In the face of this difficulty, Piketty, Saez, and Zucman (2018) choose to exclude payroll taxes and include Social Security benefits in their "pre-tax" measure of income—a departure from national accounts that requires, for the aggregates to reconcile, imputing the difference between payroll taxes and benefits to individuals. Auten and Splinter (2024a) do the opposite—including payroll taxes and excluding Social Security benefits from their pre-tax income measure. They also, separately, construct a pre-tax-plus-transfers concept of income that includes government transfer benefits. They justify this latter measure as providing a more complete picture of the total resources available for saving, consuming, and paying taxes—and as consistent with a long tradition of government measurement

[9] Post-publication revisions by Piketty, Saez, and Zucman (2018) allocate retained earnings based on their own estimates of equity wealth, but they do not fundamentally change this point, because those estimates rely on a capitalization method that leverages observed capital income.

for those purposes, even if it does not conform with notion of income from the national accounts.

This Social Security issue is connected to a fourth, and wider, problem: approaches that use national-accounts data—and that appeal to an after-tax perspective on inequality—must allocate *all* current government spending to individuals (including national defense, infrastructure, education, and so forth), because it is part of national accounts. Auten and Splinter (2024a) assign half of this spending on an equal per-capita basis, while Piketty, Saez, and Zucman (2018) assign it proportionally to other income. While one can debate the merits of both approaches (does Elon Musk benefit more from national-defense spending than you do?), there is no consensus on the right approach. Although, at a minimum, we suspect many Americans would be surprised to learn that—at least according to some researchers—defense spending primarily benefits the top of the income distribution.

Fifth, national income approaches must deal with difficulties that arise from the treatment of business investment expenses. Tax law has its own (often-changing) rules for how capital investments should be depreciated—that is, allocated over the life of the asset. These rules often do not track the real economic decline of an asset. National accounts, by contrast, attempt to estimate actual economic depreciation. Reconciling national accounts with tax data therefore requires allocating the difference in depreciation. This issue becomes especially important after the 2017 tax reform expanded the "expensing" of certain investments—that is, treating 100 percent of certain investment costs as costs in the year they were made (allowing for an immediate deduction from income), rather than depreciating (and therefore deducting) the cost of the investment over time. Piketty, Saez, and Zucman (2024) argue that Auten and Splinter (2024a) allocate to partnership owners too little of the "excess depreciation"—that is, the more generous IRS tax treatment of depreciation, relative to the national accounts. In particular, the trio argues that Auten and Splinter assign too much excess depreciation to sole proprietors. The core of the problem is the data: most excess depreciation is not directly observed on the individual tax returns. In the face of this uncertainty, Piketty, Saez, and Zucman allocate the aggregate excess depreciation of partnerships proportionally to observed partnership income. This approach is appealing in its simplicity. But it might not be correct to assume that the ownership of investment- and capital-intensive partnerships—the ones that generate the greatest excess depreciation deductions, such as utilities and real-estate companies—is well-approximated by taxpayers who have the most partnership income (predominantly finance). In their response, Auten and Splinter (2024b) discuss and defend their original allocation in detail and report the results from a new analysis based on linking the tax returns of pass-throughs (S-corporations and partnerships) and individuals (based on the work of Love 2021) to estimate the total (rather than excess) depreciation share of top 1 percent tax returns. They then make adjustments to approximate the *excess* depreciation that should be allocated to the top 1 percent of individuals. These results are similar to their original method.

Fundamentally, one approach falls back on the assumption that unobserved income is distributed similarly to some observed quantity. The other approach brings in additional microdata that provide some new distributional information—but that do not match precisely the underlying concepts. Understandably, neither approach is perfect. But the implications are surprisingly large: according to Piketty, Saez, and Zucman (2024), this one issue accounts for a 1.2 percentage point difference in the change in the top 1 percent income share between 1979 and 2019—or 30 percent of the total difference in trends between the two published papers. This issue also offers a nice illustration of how much is effectively hidden from a casual reader. A quick search for "net operating loss," "depreciation," "partnership," and "capital consumption adjustment" in the published versions of the two papers does not warn a reader of this issue. Auten and Splinter note: "We also account for other differences, such as faster depreciation in tax data than in national accounts due primarily to expensing on tax returns. See the online appendix for details." They discuss their assumptions in the appendix, asserting that they have little impact on results. Piketty, Saez, and Zucman (2018) do not comment on this particular issue at all.

A final but related issue concerns the potential re-ranking of individuals in the income distribution. When starting with income tax data, individuals are simply ranked in the income distribution according to their observed reported income. But bringing in other data requires adjustments—most importantly in this context, adjustments to individuals with low reported taxable income who are deemed to be high income when their income is "corrected." Such adjustments result in changes to the composition of the higher-income groups. This generates issues with how income that is responsible for moving individuals to top groups is treated. As a concrete example, how should we distribute the aggregate amount of income subject to tax evasion to individuals who otherwise report losses? Should we move all of them up in the income distribution uniformly, or should we move some of them up by a lot? If the latter, what assumptions should be used? Some partial microdata may be available in these and other cases—but by definition such data are not going to be comprehensive, and one needs auxiliary assumptions.

Again, this list of issues concerning the allocation of national accounts across the income distribution is not exhaustive. But it illustrates the difficulties that arise when no precise data are available. And, if income with an unknown distribution is simply allocated proportionally to observed income, then the shift to using national-accounts data would be pointless: it would just replicate the same income distribution we observe in the administrative tax data.

The turn to national income can still be defended as representing a kind of middle-ground—a middle-ground between the under-inclusive concept of income derived from tax law and the quixotic, unreachable Haig-Simons ideal. But these approaches come with tradeoffs. Narrower income concepts have better data and are easier to measure. But they are narrow, and what is included in income categories can evolve over time. Broader income concepts are, by definition, less narrow. But they are harder to measure. The harder one searches for the Haig-Simons ideal, the farther one strays from reliable data.

In practical terms, once one leaves the comfortable confines of microdata—and attempts to change the concept of income or the unit of observation—a number of problems arise. First, one needs to measure the magnitude of the additional income. Reliance on national accounts helps address this, but comes with other limitations. Second, there is the question of how the additional income is distributed relative to observed income. An attractive assumption that researchers often use is to assume that what we do not know is proportional to what we do know. As noted above, if taken to its extreme—if we assume that all unobserved income is distributed proportionally to observed income—this assumption adds no additional information. Much of the controversy over measuring income inequality stems from the choice between using this kind of easy-to-understand assumption and approaches that use imperfect auxiliary information to construct an alternative allocation. Based on various auxiliary data sources, Auten and Splinter (2024a) effectively move farther away from a proportionality assumption than Piketty, Saez, and Zucman (2018) do. Third, allowing for heterogeneity conditional on observed income changes the ordering of individuals in the distribution.

Many of the difficulties that arise in accounting for tax evasion, excess depreciation, and undistributed corporate earnings stem from these three issues. Even if one settles on the total amount of income to be allocated, variation in who this income belongs to produces uncertainty about where it falls in the observed income distribution and how the allocation of this income modifies the observed ranking.

Why It Matters and the Next Frontier

We conclude by offering a few thoughts on the implications of the issues discussed above for income measurement and inequality measurement.

First, and perhaps most obviously, there is a need for greater sensitivity to the limitations of existing data, and a need for better data. Particularly valuable is data that can link information on aggregate income with individuals. Researchers in some other countries, for example, have been constructing increasingly robust data linking firm-level income with owners: for Canada, see Wolfson et al. (2016); for France, see Bach et al. (2024); for the Netherlands, see Lejour (2024); for Chile, see Fairfield and Jorratt De Luis (2016); for Norway, see Alstadsæter et al. (2016); and in Honduras, see Del Carmen et al. (2023). This international evidence suggests that correctly allocating business income to owners can make a large difference in top income shares relative to standard imputation methods.

Some recent work on the US economy has made improvements on this front. In particular, Smith et al. (2019) link the universe of pass-through firms (firms where the entity-level income is passed along to owners as taxable income, and that do not pay a separate corporate tax) to their owners for the years 2001–2014. Love (2021) refines and extends this further for partnerships. But such work remains incomplete for some years of greatest interest—such as the 1980s, in which there were both major changes in tax law and a sudden observed increase in top income shares.

Such work also remains incomplete for other kinds of firms. For example, we do not know of any data that link the universe of C-corporations (firms that pay the separate corporate tax and that retain more income) to their individual owners.

Second, we need more attention to legal changes that affect the measurement of inequality. For example, the line between individual and firm income has shifted in the United States, stemming from legal changes in how businesses are taxed (Clarke and Kopczuk 2017, 2025). Before the Tax Reform Act of 1986, for example, one of the most common (and widely publicized) tax-planning strategies in the United States was to use C-corporations as a tax-deferral device. By retaining earnings inside a corporation, taxpayers—especially taxpayers with high marginal rates—could defer individual taxation and pay only the lower, entity-level corporate tax. But the 1986 tax reform reduced individual marginal rates and changed entity-level rules, and the incentive for individuals to use C-corporations to shelter income declined. Indeed, several hundred thousand C-corporations disappeared from the US economy in the years after 1986—replaced by pass-through entities in the form of S-corporations and, starting in the 1990s, partnerships. When firms switch in this fashion, a greater share of business income will show up on individual tax returns, and firms will change their retention behavior. How should we think about these legal changes? To what extent do changes in US income inequality reflect real economic changes versus legal changes that merely affect what we observe? Similarly, to what extent do cross-country differences reflect real differences versus differences in what national laws make measurable?

Third, we see an opportunity to appreciate the tradeoffs between different approaches to income inequality—rather than preferring a tax-based approach or a distributional national accounts approach as strictly better or worse. With either approach, we are using an imperfect measure of comprehensive income or welfare. Tax data provide a limited view of comprehensive income. But tax law defines a concept of income that can be defended as representing *society's* answer—through the democratic lawmaking process—to a hard question: "What should be counted for determining one's ability to pay for the cost of government?" The answer to this question will naturally evolve over time. We are still observing this definitional process in action today. Recent debates over taxing accrued (rather than realized) capital gains, and recent litigation over the scope of the Sixteenth Amendment, reveal that there are ongoing disagreements over exactly what constitutes, and should constitute, "income."

Fourth, we urge modesty and skepticism in the face of the profusion of income measures and inequality results. Some studies in this area produce a monolithic result with no consumer warning label: "This is the income share of the top 1 percent." Our Figure 2 reports a number of those estimates. But it would be better to think about inequality estimates as representing bounds that emerge under different assumptions. Alvaredo et al. (2024) adopt something like this approach for the case of Latin America. They note that, because no single method for measuring inequality "is fully convincing at present, we are left with (often wide) ranges, or bands, of inequality as our best summaries of inequality levels." Thinking

of individual inequality results as falling within a wide band of reasonable measures would restore a sense of perspective to this literature.

Finally, we stress that the income concept itself is nebulous and evolving. Measures with which researchers work are influenced by what "market" transactions they can observe. Likewise, governments tax what they can measure. But these things change over time. As the line between household production and market transactions shifts, "income" changes. As the retirement landscape changes—and changes the balance of various public, private, and tax-deferred savings mechanisms—when and how "income" is recognized evolves. And on and on.

One can choose a single income concept and run with it. But no single concept is perfect.

■ *For helpful comments and discussions we thank Alan Auerbach, Jerry Auten, John R. Brooks, Jim Hines, Russell Osgood, Jonathan Parker, Nina Pavcnik, Robert Pollak, Emmanuel Saez, Peter Siminski, David Splinter, Timothy Taylor, Heidi Williams, and Eric Zwick.*

References

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

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

Alvaredo, Facundo, François Bourguignon, Francisco Ferreira, and Nora Lustig. 2024. "Inequality Bands: Seventy-Five Years of Measuring Income Inequality in Latin America." World Inequality Lab Working Paper 2024/08.

Andrews, William D. 1983. "The Achilles' Heel of the Comprehensive Income Tax." In *New Directions in Federal Tax Policy for the 1980s*, edited Charles E. Walker and Mark A. Bloomfield, 278–80. Ballinger Publishing Company.

Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez. 2011. "Top Incomes in the Long Run of History." *Journal of Economic Literature* 49 (1): 3–71.

Auerbach, Alan J., Laurence J. Kotlikoff, and Darryl Koehler. 2023. "US Inequality and Fiscal Progressivity: An Intragenerational Accounting." *Journal of Political Economy* 131 (5): 1249–93.

Auten, Gerald, and David Splinter. 2024a. "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends." *Journal of Political Economy* 132 (7): 2179–227.

Auten, Gerald, and David Splinter. 2024b. "Reply to Piketty, Saez, and Zucman (2024): Income Inequality in the United States." Unpublished.

Bach, Laurent, Antoine Bozio, Arthur Guillouzouic, and Clément Malgouyres. 2024. "Do Billionaires Pay Taxes?" Unpublished.

Bittker, Boris I. 1967. "A 'Comprehensive Tax Base' as a Goal of Income Tax Reform." *Harvard Law Review* 80 (5): 925–85.

Blum, Walter J., and Harry Kalven Jr. 1952. "The Uneasy Case for Progressive Taxation." *University of Chicago Law Review* 19 (3): 417–29.

Bridgman, Benjamin, Andrew Craig, and Danit Kanal. 2022. "Accounting for Household Production in the National Accounts: An Update 1965–2020." *Survey of Current Business* 102 (2): 1–13.

Brooks, John R. 2018. "The Definitions of Income." *Tax Law Review* 71: 253–309.

Brownlee, W. Elliot. 2016. *Federal Taxation in America: A History*. 3rd ed. Cambridge University Press.

Clarke, Conor, and Wojciech Kopczuk. 2017. "Business Income and Business Taxation in the United States since the 1950s." *Tax Policy and the Economy* 31 (1): 121–59.

Clarke, Conor, and Wojciech Kopczuk. 2025. "Income Inequality and the Corporate Sector." Unpublished.

Congressional Budget Office. 2023. *The Distribution of Household Income in 2020*. Congressional Budget Office.

Dahl, Gordon B., and Michael R. Ransom. 2002. "The 10% Flat Tax: Tithing and the Definition of Income." *Economic Inquiry* 40 (1): 120–37.

Deaton, Angus. 2020. "GDP and Beyond: Summaries from the 2020 Annual Meeting of the American Economic Association." *Survey of Current Business* 100 (6): 1–5.

Del Carmen, Giselle, Santiago Garriga, Wilman Nuñez, and Thiago Scot. 2023. "Two Decades of Top Income Shares in Honduras." World Bank Policy Research Working Paper 10722.

Department of Commerce. 1934. *National Income, 1929–32*. US Government Printing Office.

Fairfield, Tasha, and Michel Jorratt De Luis. 2016. "Top Income Shares, Business Profits, and Effective Tax Rates in Contemporary Chile." *Review of Income and Wealth* 62 (S1): S120–44.

Frazis, Harley, and Jay Stewart. 2011. "How Does Household Production Affect Measured Income Inequality?" *Journal of Population Economics* 24 (1): 3–22.

Gautham, Leila, and Nancy Folbre. 2024. "Household Production and Inequality in Living Standards in the U.S., 1965–2018." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4996007.

Glogower, Ari. 2023. "A Basic Needs Baseline for Distributional Analysis." *Brigham Young University Law Review* 48 (6): 1697–768.

Graetz, Michael J., and Anne L. Alstott. 2022. *Federal Income Taxation, Principles and Policies*. 9th ed. Foundation Press.

Gruber, Jonathan. 2022. *Public Finance and Public Policy*. 7th ed. Macmillan Learning.

Haig, Robert Murray. 1921. "The Concept of Income—Economic and Legal Aspects." In *The Federal Income Tax*, edited by Robert Murray Haig, 1–28. Columbia University Press.

Hayashi, Andrew T. 2023. "The Federal Architecture of Income Inequality." Virginia Public Law and Legal Theory Research Paper 2023-80.

Joint Committee on Taxation. 2022. *Estimates of Federal Tax Expenditures for Fiscal Years 2022–2026*. JCX-22-22. Joint Committee on Taxation.

Kuznets, Simon, ed. 1953. *Shares of Upper Income Groups in Income and Savings*. NBER.

Larrimore, Jeff, Richard V. Burkhauser, Gerald Auten, and Philip Armour. 2021. "Recent Trends in US Income Distributions in Tax Record Data Using More Comprehensive Measures of Income Including Real Accrued Capital Gains." *Journal of Political Economy* 129 (5): 1319–60.

Leite, David. 2024. "The Firm as Tax Shelter: Micro Evidence and Aggregate Implications of Consumption through the Firm." Unpublished.

Lejour, Arjan. 2024. "The Economic Position of the Wealthy in the Netherlands since 2006." Unpublished.

Lockwood, Benjamin B., Charles Nathanson, and E. Glen Weyl. 2017. "Taxation and the Allocation of Talent." *Journal of Political Economy* 125 (5): 1635–82.

Love, Michael. 2021. "Where in the World Does Partnership Income Go? Evidence of a Growing Use of Tax Havens." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.3985535.

Mayerhauser, Nicole, and Marshall Reinsdorf. 2007. "Housing Services in the National Economic Accounts." Bureau of Economic Analysis, September 11. https://www.bea.gov/sites/default/files/methodologies/RIPfactsheet.pdf.

Murphy, Liam, and Thomas Nagel. 2002. *The Myth of Ownership: Taxes and Justice*. Oxford University Press.

Musgrave, R. A. 1967. "In Defense of an Income Concept." *Harvard Law Review* 81 (1): 44–62.

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

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

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2024. "Income Inequality in the United States:

A Comment." Unpublished.

Simons, Henry C. 1938. *Personal Income Taxation: The Definition of Income as a Problem of Fiscal Policy*. University of Chicago Press.

Smeeding, Timothy M., David S. Johnson, and Constance F. Citro, eds. 2024. *Creating an Integrated System of Data and Statistics on Household Income, Consumption, and Wealth: Time to Build*. National Academies of Sciences, Engineering, and Medicine. https://doi.org/10.17226/27333.

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

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

Thuronyi, Victor. 1990. "The Concept of Income." *Tax Law Review* 46 (1): 45–105.

von Kleinwächter, Friedrich Ludwig. 1896. "Das Einkommen und seine Verteilung [Income and its Distribution]." In *Hand- und Lehrbuch der Staatswissenschaften [Hand- and Textbook of Political Sciences]*, Vol. 5.

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

Zelenak, Lawrence. 2006. "Taxing Endowment." *Duke Law Journal* 55 (4): 1145–81.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 127–148

Macro Perspectives on Income Inequality

Matthieu Gomez

Inequality has emerged as a defining challenge for modern economies and a central focus of economic research over the past two decades. In this article, I highlight some key empirical insights from this literature and analyze them through the lens of economic theory. I focus on two key questions.

First, what is the appropriate notion of income to use when measuring inequality? I begin by highlighting the key differences between common measures of income, particularly in how they account for capital income. I then contrast these with the ideal notion of income from a welfare perspective—one that reflects an individual's ability to consume or save for future consumption.

Second, what are the key factors driving the recent rise in top inequality? A shift-share decomposition reveals that most of the long-term increase is not primarily driven by rising labor income inequality or shrinking labor share: rather, it is largely driven by rising capital income inequality, particularly a surge in top entrepreneurial incomes. I then apply a model of capital accumulation to quantify the role of three distinct factors behind this phenomenon: higher returns on capital (technological factors), lower costs of external financing (financial factors), and a lighter tax burden (fiscal factors).

Defining Income

Research on inequality begins with a deceptively simple question: how exactly should we define income? The choice of an income definition can dramatically influence the results, making the debates surrounding inequality more complex

■ *Matthieu Gomez is Associate Professor of Economics, Columbia University, New York City, New York. His email address is mg3901@columbia.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241435.

Table 1
Comparing Different Income Measures

Income measure	*Definition*
Panel A. Income sources	
Distributed income	= Wages + Rents + Interest received + Dividends
Factor income	= Distributed income + Retained earnings
Haig-Simons income	= Distributed income + Capital gains
Hicksian income	= Distributed income + PV[ΔDistributed income + ΔTrading profits]
Panel B. Income uses	
Distributed income	= Consumption + Asset purchases
Factor income	= Consumption + Asset purchases + Corporate investment
Haig-Simons income	= Consumption + ΔNetworth
Hicksian income	= Consumption + PV[ΔConsumption]

Note: This table compares commonly used income measures based on their sources (panel A) and uses (panel B). All income concepts satisfy the fundamental individual budget constraint: total income sources must equal total income uses. However, in a dynamic setting, each income concept differs in what qualifies as undistributed income (the portion of income that is earned but not distributed) and, consequently, what qualifies as savings (the portion of income that is earned but not consumed). See online Appendix B for mathematical definitions and further results on these income measures.

and, at times, more contentious. Take taxable income, for instance (Piketty and Saez 2003): although it is readily available from tax returns, it excludes components commonly recognized as part of economic income, including employer pension contributions or capital income received in tax-exempt accounts. Another issue with taxable income is that tax reforms can make it look like the distribution of income is changing over time—even when the distribution of economic income remains the same. Recognizing these limitations, the recent efforts by Piketty, Saez, and Zucman (2018a) and Auten and Splinter (2024) combine tax data with surveys and national accounts to focus on a more comprehensive notion of income—factor income—that aligns with national income as defined by statistical agencies.

While factor income represents an improvement, it is not a definitive solution. In this section, I highlight the conceptual limitations of three commonly used income measures in the literature: distributed income, factor income, and Haig-Simons income. I then contrast them with the ideal definition of income from a welfare perspective, known as Hicksian income. As a preview of these results, Table 1 summarizes the differences between these different income concepts based on their sources (panel A) and their uses (panel B).

Income Concepts

I first highlight the differences between three commonly used income measures—distributed income, factor income, and Haig-Simons income—with a particular focus on how they account for capital income.

Distributed Income. Distributed income is the component of economic output that is distributed to individuals. Distributed income is the sum of cash flows received

from both labor—wages, salaries, and so on—and the cash flows received from asset ownership—interest income, dividend payments, and rental income.

The concept of distributed income is simple because it focuses on the actual cash flow received by individuals. This simplicity is also its limitation, however, as current cash flows often fail to capture an individual's true economic income. To illustrate this point, consider the example of Alice, who owns stock in a firm that reinvests all of its profits rather than distributing dividends. Even though Alice receives no current cash flow from the firm, this growth in firm capital means that Alice will be able to consume more in the future, either through higher firm dividends (if she keeps holding the stock) or through higher stock prices (if she sells it).[1] This example demonstrates that distributed income fails to capture true economic income period per period. The income concepts described below will propose different ways to account for nondistributed capital income, each with its own specific adjustments and limitations.

Factor Income. Factor income includes all economic output, whether or not it is distributed to individuals. The key difference with distributed income is that it includes all firm earnings, regardless of whether they are distributed or reinvested in the firms (retained earnings). Factor income aggregates to the notion of national income as measured in national accounts (GDP minus capital depreciation plus net income from abroad), and, as such, it is the focus of Piketty, Saez, and Zucman (2018a), who aim to construct disaggregated national accounts.

One drawback of factor income, however, is that it is sensitive to changes in accounting standards. Consider, for instance, the changes in how national accounts handle firm expenses on intellectual property products—software, research and development, and artistic originals. Historically, these items were categorized as intermediate expenditures, and thus excluded from calculations of net output. However, in a series of revisions to the national income and product accounts in 1999, 2013, and 2018, the US Bureau of Economic Analysis reclassified these expenses as investments. These revisions mechanically increased the amount of retained corporate earnings, and, therefore, capital income. For instance, Koh, Santaeulàlia-Llopis, and Zheng (2020) argue that this reclassification is "responsible" for the rise in the aggregate capital share measured in national accounts. Because business ownership is heavily concentrated at the top of the income distribution, one can expect these accounting changes to affect empirical measures of income inequality as well.[2]

This sensitivity to accounting conventions highlights a more fundamental issue in using factor income to capture economic income. Consider Bob and Carol, two firm owners. Bob's firm initially generates $100 in profits and reinvests 50 percent

[1] The latter case also highlights that, even in present value terms, distributed income does not accurately capture an individual's ability to spend in the presence of trading. I will return to this point below.
[2] Note, however, that this reclassification has a more muted effect on capital income net of capital depreciation because of the increased capital depreciation associated with this new intangible capital (Rognlie 2015).

annually; this sustained investment allows Bob's firm to maintain a 5 percent annual growth in profits. Carol's firm, meanwhile, starts with $50 in profits and, despite distributing all earnings to shareholders, also achieves 5 percent annual growth because of increasing demand for its goods. Under the notion of factor income, Bob's income is $100 while Carol's income is only $50. Yet, from a financial perspective, Bob and Carol own assets that generate identical cash flows.

The general lesson is that factor income can treat assets with identical cash flows differently. It assigns higher capital income to assets whose cash flows grow through traditional investment, as recognized by national accounts, compared to those whose cash flows grow through more intangible investments (for example, brand-building) or without investment at all (for example, owning land in a rapidly developing area). Yet all these sources of cash-flow growth are fundamentally equivalent from a financial perspective. This limitation with factor income motivates another well-known measure of income, Haig-Simons income.

Haig-Simons Income. Haig-Simons income is defined as distributed income plus capital gains (whether or not they are realized)—that is, changes in the price of assets owned by individuals. Put differently, Haig-Simons's notion of capital income is equal to the return on wealth times initial wealth.[3] By supplementing distributed income with the (market-based) notion of capital gains rather than the (accounting-based) notion of retained earnings, Haig-Simons income treats all sources of asset value increases equally, whether they come from tangible investment, intangible investment, or rising productivity.

One drawback of Haig-Simons income is that not all changes in asset values reflect changes in future cash flows. In fact, a large fraction of changes in asset prices reflects variations in the interest rate at which future cash payments are discounted rather than variations in future cash flows.[4] Consider, following Cochrane (2020), a business owner Bob who consumes his firm's dividends each year. If interest rates suddenly fall, the market value of Bob's business rises, even though his future dividends—and therefore his future consumption—remain unchanged. From an economic perspective, this capital gain does not meaningfully constitute income, as it does not increase Bob's ability to consume—in effect, it is is purely a "paper gain." In sum, while factor income tends to count "too little" capital income by ignoring some sources of cash flow growth, Haig-Simons income tends to count "too much" in times of declining interest rates.

Quantitative Differences

Beyond these conceptual differences, how much do distributed income, factor income, and Haig-Simons incomes differ in practice? To answer this question, I

[3]For additional discussion of the origins of the Haig-Simons income concept, see the discussion by Clarke and Kopczuk in this symposium.

[4]For additional discussion of this point, online Appendix A clarifies the difference between these two sources of capital gains.

report these three income measures for the two largest asset classes owned by households—US corporate equity and housing—using data from the Integrated Macro Economic Accounts (Board of Governors of the Federal Reserve System 2023b). For corporate equity, I construct distributed income as corporate dividends, factor income as dividends *plus* the retained earnings of corporations, and Haig-Simons income as dividends *plus* the real change in the market value of corporate equity. For housing, I construct distributed income as rental income (including the imputed rent that homeowners would have paid to themselves if they were renting their residence). I then compute Haig-Simons income as rents *plus* the real change in the market value of housing owned by households.

The results are reported in Figure 1. The two upper panels plot the three income measures every year for the two asset classes. The Haig-Simons measures of income exhibit substantially higher year-to-year volatility, reflecting the fact that the market value of financial assets (equity or housing) fluctuates a lot over time. To focus on the low frequency trends, the two lower panels plot the average of each income measure over 20-year periods. The gap between Haig-Simons and factor income is systematically positive for housing, For corporate equity, the gap is initially negative before turning positive post 1980.[5]

What explains this positive gap between Haig-Simons and factor income for corporate equity and housing? As discussed above, a positive gap means that investors (who price assets) either expect an increase in distributed income beyond what is recorded as investment in national accounts, or anticipate lower interest rates going forward. Both mechanisms likely play a role in the data. The first mechanism—expectations of higher future distributed income—is clearly relevant for housing. Because land is fundamentally scarce, economic theory suggests that rents should grow with the economy even in the absence of investment (as in the example of Carol above).[6] A similar mechanism also seems to be at play for corporate equity as, over the past 40 years, the capital income distributed by the corporate sector has grown faster than its capital stock—an empirical pattern to which I will return in the last section of the paper. Still, the secular decline in interest rates suggests that at least some of the "excess" capital gains stem from falling discount rates too, which may not all be welfare-relevant.

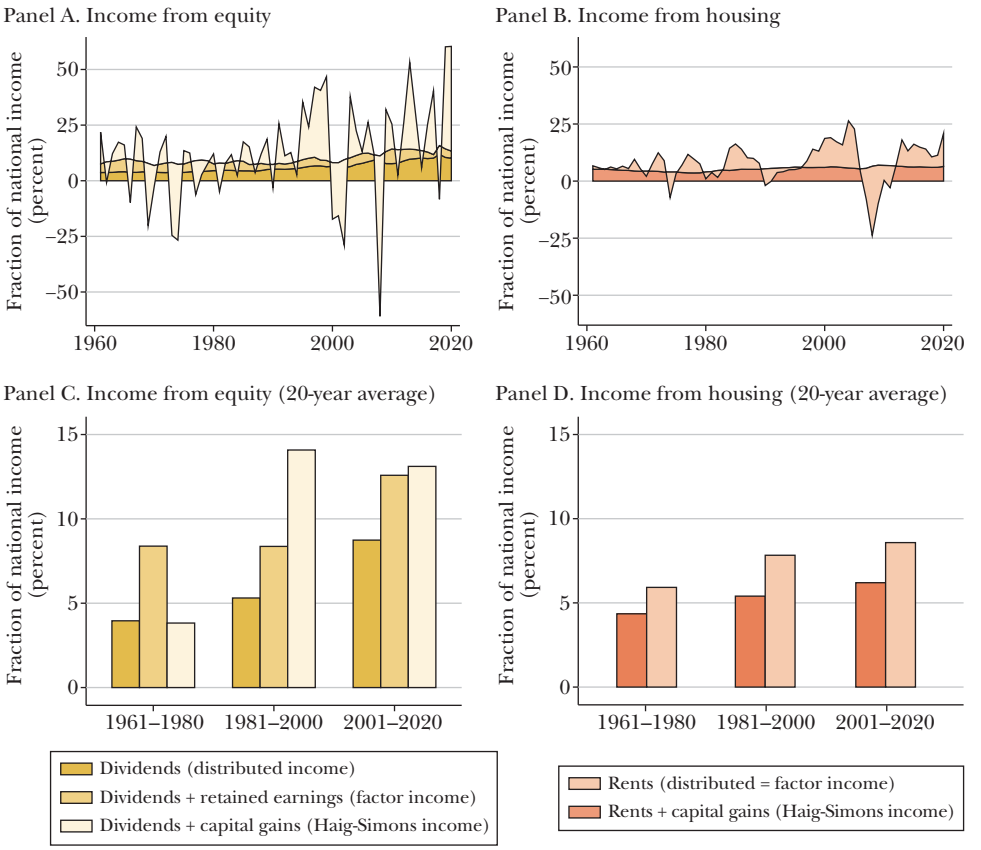
Hicksian Income

Economists have long recognized the conceptual flaws in common income measures. It was precisely this type of concerns that led Kaldor (1955) to advocate for consumption-based taxation instead. Rather than revisiting that debate, I now outline the ideal notion of income from a normative perspective.

[5]See Armour, Burkhauser, and Larrimore (2013), Robbins (2018), and Larrimore et al. (2021) for recent efforts in measuring the distribution of Haig-Simons income. Consistently with the findings in Figure 1, these studies typically find that top Haig-Simons income shares are more volatile than top factor income shares.

[6]The model in online Appendix B.2 illustrates this point.

Figure 1
Comparing Different Income Concepts for Equity and Housing



Source: Data from Board of Governors of the Federal Reserve System (2023b) and US Bureau of Labor Statistics (2023).
Note: This figure plots different notions of capital income associated with corporate equity (left panels) and housing (right panels) as a proportion of national income. For equity, I report three income concepts: distributed income (dividends), factor income (dividends plus retained earnings of corporations), and Haig-Simons income (dividends plus capital gains). For housing, I only report two income concepts: distributed income (rents) and Haig-Simons income (rent plus capital gains), as distributed income and factor income coincide.

A natural starting point is Hicks (1939), who defined income as "the maximum amount a man can spend and still be as well off at the end of the week as at the beginning." In formal terms, Hicksian income equals consumption plus the money-metric change in an individual's welfare. If someone only values consumption, this reduces to consumption plus the present value of future changes in consumption (Auerbach 1985; Sefton and Weale 2006).

Note that the concept of Hicksian income still fits with Simons's (1938) broad definition of income as "the sum of consumption and accumulation during a given

period," where the notion of accumulation is defined as the increase in the present value of future consumption. From a welfare standpoint, this is arguably the right way to think about saving, rather than the conventional approach of measuring it as the purchase of financial assets (distributed income), the purchase of financial assets and investment (factor income), or the change in financial net worth (Haig-Simons income).

As shown in Table 1, like all other notions of income, Hicksian income can also be defined in terms of its sources: it equals distributed income *plus* the present value of anticipated changes in distributed income at each horizon *plus* the present value of anticipated changes in the price of assets the individual expects to sell.[7] Like distributed income, Hicksian income emphasizes the importance of cash flows relative to other concepts such as retained earnings or capital gains. Unlike distributed income that only counts current cash flows, however, it recognizes that receiving one dollar today is equivalent to receiving $1 + r$ more dollars tomorrow, where r is the interest rate.[8] The difference between all income concepts in terms of their sources is summarized in Table 1 (panel B).

Hicksian income differs from factor and Haig-Simons income in two essential ways. First, it only counts capital gains to the extent that they affect anticipated distributed income or the price of anticipated transactions. This is the main insight of Fagereng et al. (2024), who stress that, in the absence of cash-flow growth, capital gains simply benefit future sellers and harm future buyers. Returning to the earlier example of Bob: if Bob never plans to sell, none of the capital gain constitutes Hicksian income; if Bob intends to sell his business tomorrow, the entire capital gain constitutes Hicksian income; if Bob plans to sell his business in ten years, what matters is how today's capital gain affects the expected sale price at that future date.[9]

Second, unlike factor income and Haig-Simons income, Hicksian income treats anticipated increases in dividends and wages symmetrically. To understand the rationale, compare Bob, the owner of a firm distributing \$50 in dividends today and is expected to grow 5 percent annually through reinvested earnings, and Dan, a worker earning a \$50 salary expected to grow 5 percent annually. With a 10 percent interest rate for discounting future cash flows, factor and Haig-Simons measures would assign \$100 in income to Bob but only \$50 to Dan. Yet Bob and Dan receive identical cash-flow streams: Hicksian income recognizes this equivalence by assigning \$100 in income to each.[10]

[7]For an algebraic presentation of this point, see Proposition 1 in the online Appendix.

[8]This equivalence assumes the individual's discount rate equals the interest rate. In the presence borrowing constraints, Hicksian income should instead use household-specific discount factors, see Fagereng et al. (2024).

[9]See also Dávila and Korinek (2018), Moll (2020), Glover et al. (2020), and Del Canto et al. (2023) for similar results.

[10]The present-value of a stream of cash flows that starts at C_t and grows at rate g is given by the Gordon growth formula: $C_t/(r-g)$, where r is the interest rate. Hence, Bob's Haig-Simon income—defined as the current cash flow plus the change in the value of the asset—is $C_t + gC_t/(r-g) = rC_t/(r-g)$. With $g = 5$ percent, $r = 10$ percent, and $C_t = \$50$, this yields an income of \$100. Bob's Hicksian

One might object to treating expected growth in wages and dividends symmetrically: after all, while labor income requires work, capital income does not. While I am sympathetic to this argument, such a view would contradict the very notion of income, which is based on adding up wages, dividends, and indeed income from all sources. While there is value in considering the cash flows earned from labor and those earned from asset ownership separately, the relevant question here is how to best combine these flows into a concept that accurately captures the resources available to an individual. Additionally, note that the labor-capital distinction becomes particularly blurry for active business owners, a point to which I will return in the next section.

To better understand this symmetric treatment of labor and capital income, consider the following perspective. Starting from distributed income (current cash flows), both factor and Haig-Simons income add a forward-looking component for capital—either retained earnings or capital gains. While this addition captures important information about future capital income, it creates an imbalance: capital income is effectively counted twice in a present value term, while labor income is counted only once. Hicksian income resolves this by incorporating a forward-looking component for labor income as well. This point is related to Barro (2021), who argues that the concept of national income effectively double counts capital income relative to labor income.[11]

Although Hicksian income is the ideal measure of income in a normative sense, it is difficult to quantify in practice. While factor income can be measured from observable accounting statements, and Haig-Simons income can be measured from changes in market valuations, Hicksian income requires forming expectations on future asset cash flows and asset sales—an inherently subjective exercise. Still, we can make some informed guesses about the distribution of Hicksian income relative to the distribution of factor or Haig-Simons income (for a more detailed study, see Gomez and Gouin-Bonenfant 2025a). First, because Hicksian income counts only some capital gains as income (mainly those reflecting changes in future cash flows), we can expect Hicksian capital income to fall between factor and Haig-Simons measures in periods of declining interest rates (like most of the past 30 years). Second, because Hicksian income incorporates expected wage growth, we can expect Hicksian labor income to exceed both factor and Haig-Simons measures. Because labor income is disproportionately earned by the bottom 99 percent of the distribution, this adjustment would reduce measured income inequality, to better

income—defined as current cash flow plus the present value of changes in future cash flows—coincides with Bob's Haig-Simons income in this setup (see online Appendix B for details).

[11] More generally, this discussion relates to an older literature examining the extent to which national income, as recorded in national accounts, accurately captures welfare-relevant income for the representative agent (see, for instance, Weitzman 1976; Sefton and Weale 2006; Hulten and Schreyer 2010; Barro 2021). A key message of these papers is that the two notions converge with constant technology but diverge in the presence of technological growth. As an illustration, online Appendix B.2 contrasts the representative agent's distributed income, factor income, Haig-Simons income, and Hicksian income in a standard neoclassical growth model with capital, land, and labor with growing productivity.

capture the actual differences in spending power across the population. This adjustment, however, is unlikely to affect the trend in rising inequality, as wage growth rates remain relatively stable over time.

Further Discussion

In this section, I have explored the ideal measure of income from a welfare perspective—specifically, what best reflects an individual's ability to consume now or save for the future. As a caveat, I would like to stress that this is not the same thing as the ideal measure of income for taxation. Tax policy has to grapple with a whole other set of concerns, like how easily people can shift income to avoid taxes or how taxation distorts real economic behavior. Another important distinction is that, from an optimal taxation perspective, an ideal measure of income should approximate lifetime consumption in present-value terms—as that is ultimately what matters for lifetime redistribution. In contrast, the income measures discussed above aim to measure all income earned within a given period—both consumption and savings—which implies that taxing such measures would effectively tax saved resources twice (Kaldor 1955).

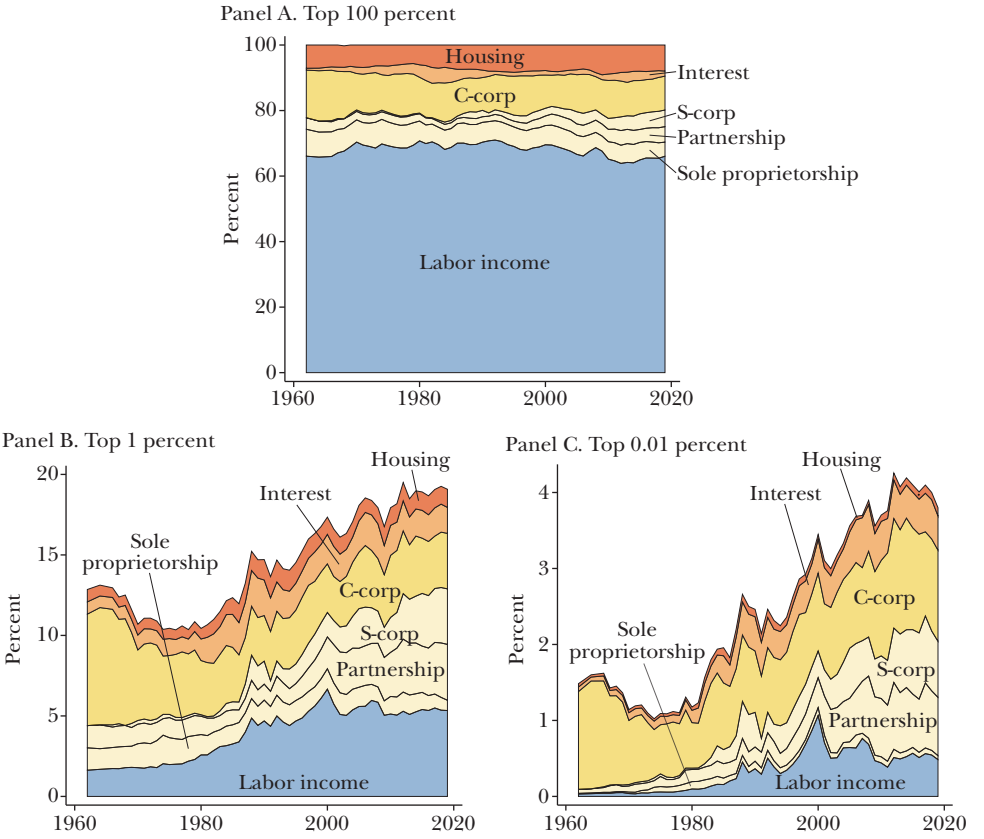
Finally, while this discussion has focused on defining income, the same conceptual challenges also emerge when defining wealth: What is the right measure—book value or market value (Bhandari and McGrattan 2021; Smith, Zidar, and Zwick 2023; Guvenen et al. 2023)? Should we treat wealth increases from declining discount rates equivalently to those from rising cash flows (Fagereng et al. 2024; Greenwald et al. 2021)? Should we include the capitalized value of future labor income (Catherine, Miller, and Sarin 2020; Greenwald et al. 2021)? At the core of these questions lies the same fundamental challenge discussed above: the extent to which a given change in wealth should be counted as economic income.

Observations on Rising Income Inequality

Regardless of the specific income concept used—be it taxable income, distributed income, factor income, or Haig-Simons income—the evidence shows rising top income inequality over the past 60 years, although researchers disagree on the exact magnitude. For example, Piketty, Saez, and Zucman (2018) find that the top 1 percent share of pre-tax factor income increased by 6 percentage points since 1960. In contrast, Auten and Splinter (2024), with different methodological choices, find a more modest increase of about 3.5 percentage points in the top 1 percent pre-tax income share, and argue that there is little change in post-tax income shares after accounting for taxes and transfers.

The trend of rising inequality is most evident at the very top: even Auten and Splinter (2024) find that post-tax income shares have grown for the top 0.1 and 0.01 percent. External sources such as rankings of the super-rich like the Forbes 400 reinforce this picture, as they show a sharp rise in both the number of billionaires and the share of aggregate wealth that they own.

Figure 2
The Composition of Income across the Distribution



Note: This figure plots the composition of the pre-tax factor income earned by top percentiles, broken down in different categories, using data from Piketty, Saez, and Zucman (2018b) with two modifications: I allocate 60 percent of pension capital income to C-corp income and the rest to interest income instead of treating it as a separate source of capital income, and I offset mortgage payments against interest income rather than housing income.

Composition of Top Incomes

The empirical literature does not just tell us how much top income shares have grown—it also sheds light on where that income comes from. Panel A of Figure 2 decomposes US national income (that is, aggregate factor income) into its four main components over time: labor compensation (68 percent of total since 1962), business income (22 percent), interest income (2 percent), and rental income from housing (8 percent).

Two trends are apparent. First, the share of national income that takes the form of labor compensation (wages and employer pension contributions) has remained relatively stable over the sample period, rising slightly from 1960 to 1980 before declining modestly through 2020. This stability stands in contrast to

well-documented decline in the "gross" labor since 1960 (Karabarbounis 2024). The difference between the gross labor share (labor income as a fraction of GDP) and the "net" labor share (labor income as a fraction of national income) arises from a rise in the depreciation rate of capital, reflecting a shift toward higher-depreciation assets such as software and computers (Rognlie 2015).[12]

Second, the composition of business income has shifted significantly. Following the 1986 Tax Reform Act, which raised corporate rates while lowering personal rates, businesses increasingly moved from C-corporations, where owners are taxed separately from the business, to pass-through entities, such as S-corporations, partnerships, and sole proprietorships, whose income flows directly to households for tax purposes. C-corporations tend to be large firms with a diffuse ownership, while pass-through businesses tend to be small or mid-market firms with concentrated, active owners.

Panels B and C of Figure 2 plots the composition of income accruing to the top 1 percent and top 0.01 percent (data from Piketty, Saez, and Zucman 2018b). Unlike the relatively stable composition of national income, the sources of top income have shifted markedly, with a rising share of labor compensation as well as income from pass-through businesses. In 1960, the top 1 percent's income was primarily derived from C-corporations; by 2019, it was evenly split between labor income, pass-through business income, and other forms of capital income.

The rising importance of pass-through businesses poses significant measurement challenges. A substantial portion of factor income from these businesses is untaxed due to generous depreciation provisions in the tax code, and, as a result, does not appear on individual tax returns. How to allocate this untaxed income property to individual earners remains a key issue due to the lack of microdata linking businesses to their owners. In fact, current debates over the level of inequality largely come down to disagreements over how to properly allocate the portion of national income and wealth earned through pass-through businesses (Auten and Splinter 2024; Smith et al. 2019; Smith, Zidar, and Zwick 2023; Piketty, Saez, and Zucman 2024).

Shift-Share Decomposition

To analyze the rise in top income shares, I now use a shift-share approach that separates the impact of three forces: rising labor income inequality, rising capital income inequality, and a shrinking labor share. For any given top percentile p, the share of total income going to that group can be expressed as a weighted average of their share of total labor income and their share of total capital income:

$$\begin{aligned} \textit{Income share}(p) \;=\; & \textit{LS} \times \textit{Labor income share}(p) \\ & + (1 - \textit{LS}) \times \textit{Capital income share}(p), \end{aligned}$$

[12] Technically, both labor share measures should also include the fraction of income from noncorporate businesses (partnerships and sole proprietorships) that accrues to labor. In practice, this adjustment has minimal impact because analyses typically assume the labor-capital split in the noncorporate sector mirrors that of the corporate sector.

where *LS* denotes the economy's overall labor share. Taking the difference of this equation over time gives a way to decompose the rise in top income shares:[13]

$$\begin{aligned}\Delta \textit{Income share}(p) \;=\; & \textit{LS} \times \Delta \textit{Labor income share}(p) \\ & + (1 - \textit{LS}) \times \Delta \textit{Capital income share}(p) \\ & + [\textit{Labor income share}(p) - \textit{Capital income share}(p)] \times \Delta \textit{LS}.\end{aligned}$$

Each term has a clear interpretation. The first term captures how changes in labor income inequality affects top income shares. The second captures how changes in capital income inequality affects top income shares. The third term captures the impact of changes in the economy's overall labor share, because capital income is more unequally distributed than labor income. For related decompositions, see Meade (2013), Moll, Rachel, and Restrepo (2022), and Irie (2024).

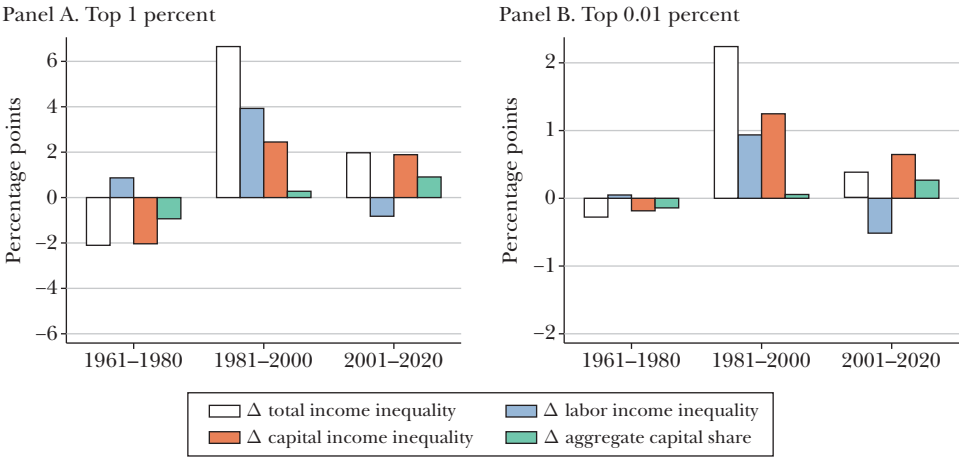
Classifying income from pass-through businesses as either labor or capital is challenging. Part of the challenge is empirical: noncorporate businesses do not separate owner wages from profits on tax returns. Even S-corporation owners, who are required make this distinction, have tax incentives to classify labor income as profits (Smith et al. 2022). The difficulty is also conceptual—defining labor and capital income in these cases is not straightforward. Consider profits from organizational capital (for example, know-how or customer bases). These profits are not purely labor income, as they compensate for effort exerted in the past rather than current work. Yet, they are not purely capital income either, because this kind of capital is effectively embodied in the owners themselves (Jones and Kim 2018; Smith et al. 2019; Bhandari and McGrattan 2021; Eisfeldt, Falato, Xiaolan 2023; Crouzet et al. 2022). Following Saez and Zucman (2020), I use a 75/25 labor/capital split for small pass-through businesses and a 25/75 split for the largest ones. However, the specific allocation has little impact on the decomposition—the results remain largely unchanged when using the classification from Smith et al. (2019), which tends to use a higher labor/capital split.

Figure 3 presents the results of the accounting framework for the top 1 percent and top 0.01 percent. To visualize the findings, I aggregate the results of the annual decomposition over three 20-year periods: 1960–1980, 1980–2000, and 2000–2020. The first key takeaway is that the decline in the aggregate labor share had minimal impact on top income shares. In other words, rising inequality is almost entirely driven by growing disparities within labor and capital income—not by a shift from labor to capital.

Second, most fluctuations in top income shares are driven by fluctuations in capital income inequality rather than fluctuations in labor income inequality. The main exception is the period between 1981 and 2000, when a significant portion of

[13] The Δ notation denotes the difference between the two periods. See online Appendix C for a formal derivation of the equation in the text.

Figure 3
A Shift-Share Decomposition of the Rise in Top Income Inequality



Source: Data from Piketty, Saez, and Zucman (2018b).
Note: This figure reports the results of using a shift-share approach to decompose the overall change in top factor (pretax) income shares into three components: rising labor income inequality, rising capital income inequality, and declining labor share. I implement this decomposition yearly and aggregate the results over three periods: 1962–1982, 1982–2002, and 2002–2020. See online Appendix C for additional details and results.

the rise in top income shares was driven by a rise in labor income inequality. This pattern aligns with research documenting the surge in executives' pay and, more broadly, the growing returns to talent during that time (see, among many others, Levy and Murnane 1992; Jones and Kim 2013; Edmans, Gabaix, and Jenter 2017).

Drivers of Rising Income Inequality

I now study the proximate causes behind the rise in top inequality. A frequently used method in the literature is to adopt a *forward-looking* approach: start with top earners at the beginning of the period and follow their income trajectories over time. However, this approach is fundamentally flawed because the composition of individuals in the top percentiles is constantly changing—yesterday's highest earners are not necessarily today's. For instance, based on the IRS public use panel data from 1979 to 1990, Gomez (2023) finds that the average income growth of existing individuals in the top percentiles was zero; instead, the entire increase in top income shares was driven by composition changes—individuals with high income growth entering the top percentiles and displacing previous top earners. This pattern persists even when income is smoothed using a three-year moving average to reduce year-to-year fluctuations.

Given this constant turnover, a more informative approach is to adopt a *backward-looking* perspective—examining how the lifetime trajectories of today's top earners compare with those of their predecessors (Gomez and Gouin-Bonenfant 2024; Ozkan et al. 2023; Gomez 2024). Pragmatically, this means asking: How does the lifetime income trajectory of today's top earners compare to that of previous generations? What has changed in the way income is accumulated?

A Simple Model of Capital Accumulation

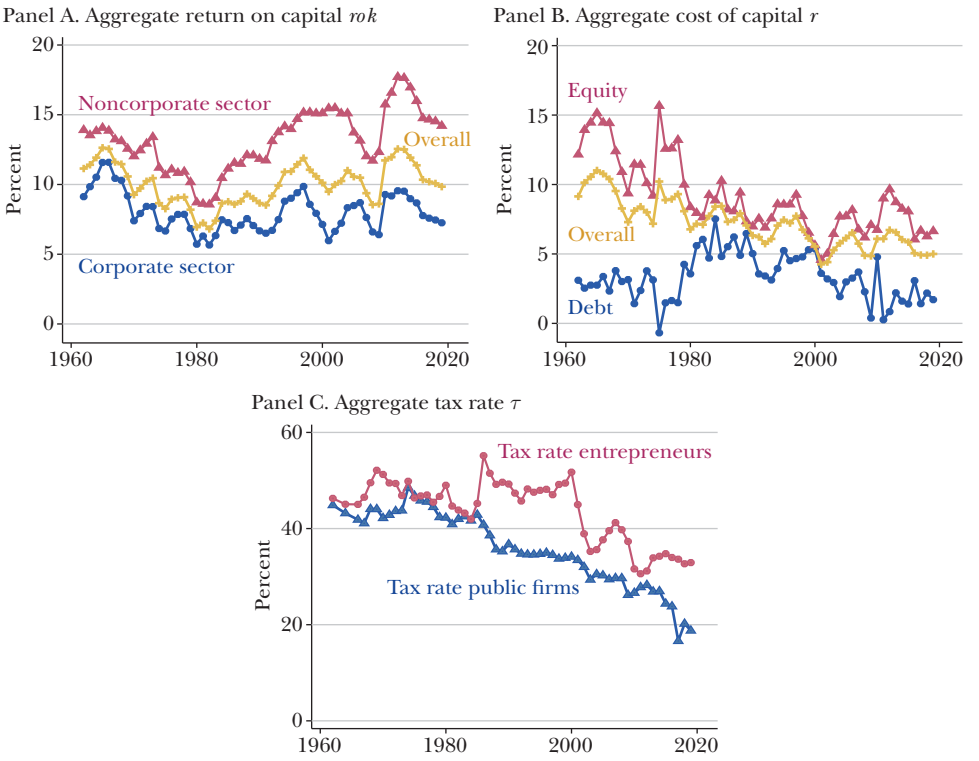
Given its central role in the trend of rising top income shares (show earlier in Figure 3), I focus on analyzing the rise in capital income inequality, and, more specifically, the rise in the income accruing to top entrepreneurs. Here, and throughout the rest of the section, I will use the term entrepreneurs to refer to individuals whose income primarily derives from business ownership—whether of public or private firms, and whether they are actively involved in management or are passive owners. I use a simple model of capital accumulation. Each period, entrepreneurs can borrow external funds at an interest rate r. They operate a project that produces a return on capital (profit) rok_i per dollar invested in the project. The profits from the projects are then used to pay interest expenses and taxes, to consume, or to invest in new capital. This model leads to the following accounting equation for entrepreneurs' capital accumulation:

$$\textit{rate of capital accumulation}_i = (1 - \tau)[r + \lambda(rok_i - r)] - \textit{consumption rate}_i,$$

where λ denotes the entrepreneur's leverage—the amount of capital operated by the entrepreneur divided by the amount of capital they own—and where τ denotes the effective tax rate faced by the entrepreneur. Note that the return on capital, rok_i, is conceptually different from the cost of capital, r. The return on capital, rok_i, represents the entrepreneur's profit per unit of capital invested in the firm—it is a physical return, largely determined by the entrepreneur's production technology—while the cost of capital, r, represents how much it costs for the entrepreneur to borrow external funds and is a financial rate of return, determined by the supply and demand for funds in financial markets. In the neoclassical model with linear return to scale, entrepreneurs borrow capital until the marginal product of capital equals the interest rate; and so $rok_i = r$. In practice, however, the two objects typically differ at the individual level because of market power, decreasing return to scale, or adjustment costs in installing capital (for example, Hayashi 1982).

This capital accumulation equation highlights three key determinants of capital income inequality identified in the literature: the return on capital, rok_i (Moll, Rachel, and Restrepo 2022), the cost of capital, r (Gomez and Gouin-Bonenfant 2024; İmrohoroğlu and Zhao 2022), and the average tax rate, τ (Hubmer, Krusell, and Smith 2021; Lee et al. 2021). I now discuss the long-term trends in each of these factors.

Figure 4
Economic Determinants of Entrepreneurs' Growth Rates



Note: This figure plots the evolution of key economic objects that affect the entrepreneurs' rate of capital accumulations. Panel A of Figure 4 plots the return on capital, computed as the ratio of a sector's net operating surplus to its capital (data from Board of Governors of the Federal Reserve System 2023b). Panel B of Figure 4 plots the cost of capital of the corporate sector, calculated as the ratio of net operating surplus minus net investment to the market value of liabilities, plus net investment (for additional details, see Gomez and Gouin-Bonenfant 2024). The cost of debt is calculated as the weighted average of the yield of AAA bonds and of the bank prime loan rate, adjusted for inflation (data from Moody's 2023; Board of Governors of the Federal Reserve System 2023a; US Bureau of Labor Statistics 2023). The cost of equity is calculated using the Modigliani-Miller formula $r_D + \lambda(r_A - r_D)$ where λ denotes the leverage of the corporate sector, r_A the overall cost of capital, and r_D the cost of debt. Finally, panel C of Figure 4 computes corporate taxes paid, as a percentage of net income, for firms in Compustat as well as total taxes paid, as a percentage of pre-tax factor income, for entrepreneurs in the DINA microdata (data from SP Global Market Intelligence 2025; Piketty, Saez, and Zucman 2018b).

Return on Capital

I construct the aggregate return on capital in the US corporate and noncorporate sector by dividing net operating surplus by the book value of capital, using data from Board of Governors of the Federal Reserve System (2023b). The results, plotted in Panel A of Figure 4, indicate that the return on capital exhibits a U-shape pattern,

declining between 1960 and 1980 and rising thereafter. This pattern closely aligns with the dynamics of top income shares over the same period.[14]

I also present separate estimates of the return on capital for the corporate and noncorporate sectors. I find that the recent increase in the return on capital is largely driven by firms in the noncorporate sector—partnerships and sole proprietorships—whose ownership is concentrated at the top. While intriguing, this difference is difficult to interpret due to significant composition changes over time—for example, a firm that previously filed as an S-corporation may now file as a partnership.

While this empirical evidence focuses on changes in the *average* return on capital, the more relevant measure in our context would be the changes in the return on capital for the right tail of entrepreneurs; that is, those who make it to top percentiles. This is more difficult to measure due to the lack of microdata linking individuals to private firms as well as difficulties in measuring capital (tangible and intangible) for individual firms. Still, some evidence suggests that the increase in the return on productive assets is concentrated in a small number of fast growing firms (for example, Andrews, Criscuolo, and Gal 2016; Autor et al. 2020), which are typically owned by top entrepreneurs.

What explains the rise in the return on capital? One possible explanation is higher productivity, thanks in part to technological advances like computers and automation (for example, Moll, Rachel, and Restrepo 2022). Another explanation is that firms are gaining more market power—both in the product and labor markets (for example, De Loecker, Eeckhout, and Unger 2020; Boar and Midrigan 2024).

Cost of Capital

I now examine the evolution of the entrepreneurs' cost of capital. As shown in the capital accumulation equation, entrepreneurs benefit from a decrease in the interest rate, r, provided they are net borrowers (that is, the leverage $\lambda > 1$). Panel B of Figure 4 plots the average cost of capital for the US corporate sector, measured as the payout yield of the corporate sector plus the growth in capital. The figure illustrates a steady decline in the average cost of funding since 1960.

Entrepreneurs raise external financing through both debt and equity. To differentiate between the two, panel B of Figure 4 presents separate estimates of the cost of capital for debt and equity financing. The cost of debt is computed as a weighted average of the bank prime loan rate and the yield on AAA bonds adjusted for inflation, while the cost of equity is calculated by combining the cost of capital with the cost of debt using the usual Modigliani-Miller formula (as explained in the notes under Figure 4). The figure shows that the cost of debt is initially low, spikes in the 1980s, and then declines. In contrast, the cost of equity steeply declines over time. Hence, lower cost of capital benefited both debt and equity issuers.

[14] For more detail on this point, see online Appendix Figure A1.

What explains the decline in the cost of capital over time? The literature does not provide a definitive answer, but proposed explanations include the slowdown in average productivity growth, increased foreign demand for US assets (the "global savings glut"), population aging (Auclert et al. 2021), and, possibly, rising inequality itself (Mian, Straub, and Sufi 2021; Gomez 2024 forthcoming).

Tax Rate

I now examine the evolution of the average tax rate paid by entrepreneurs. I start by computing the effective tax rate for public firms, which I define as corporate taxes paid divided by net income (net operating surplus minus interest payments). As shown in Figure 4, panel C, this effective rate has declined from 40 percent to 20 percent since 1962. This is consistent with the overall decline in the statutory corporate tax rate during the period, which decreased from 53 percent in 1962 to 21 percent under the Tax Cuts and Jobs Act in 2017.

The decline in the effective corporate tax rate may not fully capture the effective tax rate of entrepreneurs, as a substantial fraction of them own pass-through businesses, which are not subject to the corporate tax. Therefore, I also estimate the average tax rate of entrepreneurs using microdata from the Distributional National Accounts (Piketty, Saez, and Zucman 2018a). I define an individual as an entrepreneur if factor income from businesses (corporate and noncorporate) accounts for more than half of their total revenue—this group represents roughly 15 percent of the population. Panel C of Figure 4 reports that the average tax rate of entrepreneurs under this computation follows a similar downward trend, from 50 percent to 30 percent between 1962 and 2019. This decline in the effective tax rate reflects both lower marginal tax rates and greater opportunities for entrepreneurs to reduce their taxable income, for instance, through generous depreciation allowances. Note, however, that there is substantial disagreement about whether tax rates have effectively declined for this segment of the population. In particular, Auten and Splinter (2024) suggest that tax rates have remained stable for this group.

Quantification

I now perform a back-of-the-envelope calculation to quantify the mechanical effect of these long-term trends on top income inequality. Figure 4 reports that, since 1980, the average return on capital has increased by 2 percentage points, the cost of capital has decreased by 3 percentage points, and the average tax rate has declined by 10 percentage points. Plugging these numbers into the capital accumulation equation above suggests that, together, these trends raised the capital growth rate for top entrepreneurs by 7.5 percentage points (3 percentage points due to changes in rok, 1.5 percentage points due to changes in r, and 3 percentage points due to changes in τ)—for this computation, I assume an average leverage of $\lambda = 1.5$ and an average capital accumulation rate of 30 percentage points for entrepreneurs who make it to the top percentiles, consistently with Gomez and Gouin-Bonenfant (2024). Assuming that the average entrepreneur in the top percentiles has been

operating a firm for around 20 years, this translates into a 1.5 log point increase in their capital holdings relative to their counterparts in 1980. Figure 1 shows that the income share of the top 0.01 percent has quadrupled since 1980—a 1.4 log point increase—which is of a similar order of magnitude. Thus, this calculation suggests that the combined effects of changes in returns to capital, interest rates, and taxes can quantitatively account for the rise in top income shares.[15]

In this exercise, I have taken the changes in the return on capital, the interest rate, and the tax rate as given. From an economic perspective, however, the observed divergence between these three objects raises a puzzle: when the after-tax return on capital is high relative to the interest rate, as it is now, it should incentivize more people to start businesses and encourage existing entrepreneurs to invest more. In turn, this surge in entrepreneurial investment should push down the return on capital and, ultimately, decrease inequality.

A key question is why this equilibrium effect has not yet occurred (or, in the terminology of Barkai 2020, what explains the rise in "pure profits"). One explanation is that entrants now face increased barriers to competition with established firms due to changes in knowledge diffusion (Akcigit and Ates 2023), regulatory capture (Gutiérrez and Philippon 2019), or demographics (Karahan, Pugsley, and Şahin 2024). Another possibility is that the aggregate supply of entrepreneurial talent is less elastic than commonly assumed (Gomez and Gouin-Bonenfant 2025b).

In the absence of such equilibrium effects, government policies could act as an additional stabilizing force on inequality. In practice, however, this stabilizing role of government policies has remained limited in the United States, possibly because of prevailing beliefs that high inequality reflects a meritocratic process (Mijs 2021), the challenge of enacting nondistortionary taxes on the wealthy (Bastani and Waldenström 2020), or the influence of entrenched elites on policymaking (Glaeser, Scheinkman, and Shleifer 2003).

References

Akcigit, Ufuk, and Sina T. Ates. 2023. "What Happened to US Business Dynamism?" *Journal of Political Economy* 131 (8): 2059–2124.

Andrews, Dan, Chiara Criscuolo, and Peter N. Gal. 2016. "The Best versus the Rest: The Global Productivity Slowdown, Divergence across Firms and the Role of Public Policy." OECD Productivity Working Paper 2016-05.

Armour, Philip, Richard V. Burkhauser, and Jeff Larrimore. 2013. "Levels and Trends in United States Income and Its Distribution: A Crosswalk from Market Income towards a Comprehensive Haig-Simons Income Approach." NBER Working Paper 19110.

[15] Naturally, this is a simplified exercise, and I discuss some caveats in online Appendix D.

Auclert, Adrien, Hannes Malmberg, Frédéric Martenet, and Matthew Rognlie. 2021. "Demographics, Wealth, and Global Imbalances in the Twenty-First Century." NBER Working Paper 29161.

Auerbach, Alan J. 1985. "Saving in the US: Some Conceptual Issues." In *The Level and Composition of Household Saving*, edited by Patric H. Hendershott. Ballinger Pub. Co.

Auten, Gerald, and David Splinter. 2024. "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends." *Journal of Political Economy* 132 (7): 2179–2227.

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

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

Barro, Robert J. 2021. "Double Counting of Investment." *Economic Journal* 131 (638): 2333–56.

Bastani, Spencer, and Daniel Waldenström. 2020. "How Should Capital Be Taxed?" *Journal of Economic Surveys* 34 (4): 812–46.

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

Boar, Corina, and Virgiliu Midrigan. 2024. "Markups and Inequality." *Review of Economic Studies*. https://doi.org/10.1093/restud/rdae103.

Board of Governors of the Federal Reserve System. 2023a. *Bank Prime Loan Rate DPRIME*. Distributed by FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DPRIME (accessed November 9, 2023).

Board of Governors of the Federal Reserve System. 2023b. *Integrated Macroeconomic Accounts for the United States*. 2023. https://www.federalreserve.gov/apps/fof/FOFTables.aspx#integrated (accessed November 11, 2023).

Catherine, Sylvain, Max Miller, and Natasha Sarin. 2020. "Social Security and Trends in Inequality." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.3546668.

Cochrane, John. 2020. "Wealth and Taxes, Part II." The Grumpy Economist (blog), January 3. https://johnhcochrane.blogspot.com/202%1/wealth-and-taxes-part-ii.html.

Crouzet, Nicolas, Janice C. Eberly, Andrea L. Eisfeldt, and Dimitris Papanikolaou. 2022. "The Economics of Intangible Capital." *Journal of Economic Perspectives* 36 (3): 29–52.

Dávila, Eduardo, and Anton Korinek. 2018. "Pecuniary Externalities in Economies with Financial Frictions." *Review of Economic Studies* 85 (1): 352–95.

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

Del Canto, Felipe N., John R. Grigsby, Eric Qian, and Conor Walsh. 2023. "Are Inflationary Shocks Regressive? A Feasible Set Approach." NBER Working Paper 31124.

Edmans, Alex, Xavier Gabaix, and Dirk Jenter. 2017. "Executive Compensation: A Survey of Theory and Evidence." In *The Handbook of the Economics of Corporate Governance*, Vol. 1, edited by Benjamin E. Hermalin and Michael S. Weisbach, 383–539. Elsevier.

Eisfeldt, Andrea L., Antonio Falato, and Mindy Z. Xiaolan. 2023. "Human Capitalists." In *NBER Macroeconomics Annual*, Vol. 37, 1–61. University of Chicago Press.

Fagereng, Andreas, Matthieu Gomez, Émilien Gouin-Bonenfant, Martin Holm, Benjamin Moll, and Gisle Natvik. 2024. "Asset-Price Redistribution." World Inequality Lab Working Paper 2024/14.

Glaeser, Edward, Jose Scheinkman, and Andrei Shleifer. 2003. "The Injustice of Inequality." *Journal of Monetary Economics* 50 (1): 199–222.

Glover, Andrew, Jonathan Heathcote, Dirk Krueger, and José-Víctor Ríos-Rull. 2020. "Intergenerational Redistribution in the Great Recession." *Journal of Political Economy* 128 (10): 3730–78.

Gomez, Matthieu. 2023. "Decomposing the Growth of Top Wealth Shares." *Econometrica* 91 (3): 979–1024.

Gomez, Matthieu. 2024. "Counterfactual Wealth Distributions." Unpublished.

Gomez, Matthieu. 2025. *Data and Code for: "Macro Perspectives on Income Inequality."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E222942V1.

Gomez, Matthieu. Forthcoming. "Wealth Inequality and Asset Prices." *Review of Economic Studies*. https://doi.org/10.1093/restud/rdaf008.

Gomez, Matthieu, and Émilien Gouin-Bonenfant. 2024. "Wealth Inequality in a Low Rate Environment." *Econometrica* 92 (1): 201–46.

Gomez, Matthieu, and Émilien Gouin-Bonenfant. 2025a. "Welfare Inequality." Unpublished.

Gomez, Matthieu, and Émilien Gouin-Bonenfant. 2025b. "Inelastic Capital in Intangible Economies." Unpublished.

Greenwald, Daniel L., Matteo Leombroni, Hanno Lustig, and Stijn Van Nieuwerburgh. 2021. "Financial and Total Wealth Inequality with Declining Interest Rates." NBER Working Paper 28613.

Gutiérrez, Germán, and Thomas Philippon. 2019. "The Failure of Free Entry." NBER Working Paper 26001.

Guvenen, Fatih, Gueorgui Kambourov, Burhan Kuruscu, Sergio Ocampo, and Daphne Chen. 2023. "Use It or Lose It: Efficiency and Redistributional Effects of Wealth Taxation." *Quarterly Journal of Economics* 138 (2): 835–94.

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

Hicks, John R. 1939. "The Foundations of Welfare Economics." *Economic Journal* 49 (196): 696–712.

Hubmer, Joachim, Per Krusell, and Anthony A. Smith Jr. 2021. "Sources of US Wealth Inequality: Past, Present, and Future." In *NBER Macroeconomics Annual*, Vol. 35, 391–455. University of Chicago Press.

Hulten, Charles R., and Paul Schreyer. 2010. "GDP, Technical Change, and the Measurement of Net Income: The Weitzman Model Revisited." NBER Working Paper 16010.

İmrohoroğlu, Ayşe and Kai Zhao. 2022. "Rising Wealth Inequality: Intergenerational Links, Entrepreneurship, and the Decline in Interest Rate." *Journal of Monetary Economics* 127: 86–104.

Irie, Magnus. 2024. "Wealth Inequality and Changing Asset Valuations in the Distributional National Accounts." Unpublished.

Jones, Charles I., and Jihee Kim. 2013. "Exploring the Dynamics of Top Income Inequality." Unpublished.

Jones, Charles I., and Jihee Kim. 2018. "A Schumpeterian Model of Top Income Inequality." *Journal of Political Economy* 126 (5): 1785–1826.

Kaldor, Nicholas. 1955. *An Expenditure Tax*. Routledge.

Karabarbounis, Loukas. 2024. "Perspectives on the Labor Share." *Journal of Economic Perspectives* 38 (2): 107–36.

Karahan, Fatih, Benjamin Pugsley, and Ayşegül Şahin. 2024. "Demographic Origins of the Start-Up Deficit." *American Economic Review* 114 (7): 1986–2023.

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

Larrimore, Jeff, Richard V. Burkhauser, Gerald Auten, and Philip Armour. 2021. "Recent Trends in US Income Distributions in Tax Record Data Using More Comprehensive Measures of Income Including Real Accrued Capital Gains." *Journal of Political Economy* 129 (5): 1319–60.

Lee, Ji Hyung, Yuya Sasaki, Alexis Akira Toda, and Yulong Wang. 2021. "Fixed-k Tail Regression: New Evidence on Tax and Wealth Inequality from Forbes 400." Preprint, arXiv. https://doi.org/10.48550/arXiv.2105.10007.

Levy, Frank, and Richard J. Murnane. 1992. "US Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations." *Journal of Economic Literature* 30 (3): 1333–81.

Meade, James E. 2013. *Efficiency, Equality and the Ownership of Property*. Routledge.

Mian, Atif, Ludwig Straub, and Amir Sufi. 2021. "Indebted Demand." *Quarterly Journal of Economics* 136 (4): 2243–2307.

Mijs, Jonathan J. B. 2021. "The Paradox of Inequality: Income Inequality and Belief in Meritocracy Go Hand in Hand." *Socio-Economic Review* 19 (1): 7–35.

Moll, Benjamin. 2020. "Comment on 'Sources of U.S. Wealth Inequality: Past, Present, and Future.'" In *NBER Macroeconomics Annual*, Vol. 35. University of Chicago Press.

Moll, Benjamin, Lukasz Rachel, and Pascual Restrepo. 2022. "Uneven Growth: Automation's Impact on Income and Wealth Inequality." *Econometrica* 90 (6): 2645–83.

Moody's. 2023. *Moody's Seasoned Aaa Corporate Bond Yield (AAA)*. Distributed by FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/AAA (accessed November 11, 2023).

Ozkan, Serdar, Joachim Hubmer, Sergio Salgado, and Elin Halvorsen. 2023. "Why Are the Wealthiest So Wealthy? A Longitudinal Empirical Investigation." CESifo Working Paper 10324.

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

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

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2018b. *Data for: "Distributional National Accounts: Methods and Estimates for the United States."* University of Chicago Press. https://gabriel-zucman.eu/usdina/ (accessed December 1 2024).

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2024. "Income Inequality in the United States: A Comment." Unpublished.

Robbins, Jacob A. 2018. "Capital Gains and the Distribution of Income in the United States." Unpublished.

Rognlie, Matthew. 2015. "Deciphering the Fall and Rise in the Net Capital Share: Accumulation or Scarcity?" *Brookings Papers on Economic Activity* 46 (1): 1–69.

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

Saez, Emmanuel, and Gabriel Zucman. 2020. "Trends in US Income and Wealth Inequality: Revising after the Revisionists." NBER Working Paper 27921.

Sefton, James A., and Martin R. Weale. 2006. "The Concept of Income in a General Equilibrium." *Review of Economic Studies* 73 (1): 219–49.

Simons, Henry C. 1938. *Personal Income Taxation: The Definition of Income as a Problem of Fiscal Policy*. University of Chicago Press.

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

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2022. "The Rise of Pass-Throughs and the Decline of the Labor Share." *American Economic Review: Insights* 4 (3): 323–40.

Smith, Matthew, Owen Zidar, and Eric Zwick. 2023. "Top Wealth in America: New Estimates under Heterogeneous Returns." *Quarterly Journal of Economics* 138 (1): 515–73.

SP Global Market Intelligence. 2025. *Compustat Annual Data*. Distributed by WRDS. https://wrds-web.wharton.upenn.edu/ (accessed March 14, 2025).

US Bureau of Labor Statistics. 2023. *Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCSL)*. Distributed by FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/CPIAUCSL (accessed September 11, 2023).

Weitzman, Martin L. 1976. "On the Welfare Significance of National Product in a Dynamic Economy." *Quarterly Journal of Economics* 90 (1): 156–62.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 149–170

Public Finance Implications of Economic Inequality

Alan J. Auerbach

As an undergraduate in the 1970s, I learned that economic inequality in the United States, though considerable early in the twentieth century, had fallen substantially over time, and that in the decades after World War II the lower degree of inequality and its relative stability had made inequality a less central economic and political issue than it had once been. I turned my attention elsewhere.

Since then, however, many studies have suggested a strong and continuing increase in inequality, and the issue has certainly come to receive more attention from economists. Although traditional concerns about inequality have often concentrated on the well-being of the least affluent, for whom small changes in resources can have significant effects on well-being and thus on measures of social welfare, much of the recent literature has focused on the other end of the income distribution. Of particular note is the influential work of Piketty and Saez (2003) and Piketty, Saez, and Zucman (2018), who argue that income inequality has risen sharply over roughly the past 45 years, particularly if one focuses on those in the top 1 percent or in the top fraction of 1 percent. Saez and Zucman (2016) deliver the same message about the trend in US wealth inequality. But measuring inequality using available data is a challenging task, and there have been many substantive disagreements over which assumptions are most appropriate for filling in gaps in the data, notably Smith, Zidar, and Zwick (2023) regarding wealth inequality and Auten and Splinter (2024) regarding income inequality. The paper in this symposium by Clarke and Kopczuk focuses on the important questions of income

■ *Alan J. Auerbach is Robert D. Burch Professor of Economics and Law, University of California, Berkeley, Berkeley, California. His email address is auerbach@econ.berkeley.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241421.

measurement, and I will not attempt to do so here. However, it is worth noting that even under the alternative assumptions favored by critics, an upward trend remains in the inequality of both wealth and a comprehensive measure of income before taxes and transfer payments.

In this essay, without resolving disputes about income and wealth measurement, we can move on to questions about the implications of rising inequality for the theory and practice of public finance. The first two sections of this paper address fundamental reasons why the distribution of income or wealth on an annual basis before taxes and transfers[1] offers insufficient information: (1) it does not tell us what resources are actually available to households for consumption; and (2) in providing a snapshot of the resources available to individuals of different ages at a given moment in time, without controlling for life-cycle related differences or income dynamics, it can provide a misleading estimate of the underlying degree of inequality. Addressing these topics requires looking into the incidence and labelling of taxes, the value that people place on government spending, and looking at how taxes and benefits of programs like Social Security should be accounted for in the analysis of inequality. The following section of the paper considers the implications of high and perhaps rising economic inequality for the design of government policy: top marginal tax rates, phase-outs of government policies for those with higher incomes, the political economy of inequality, and other subjects.

Measuring the Distributional Effects of Government Policies

If we are concerned about the well-being of individuals, traditional economic analysis would suggest we consider only the resources available to them, not the source of those resources—that is, disregarding whether the source is market activity or the actions of government. Thus, we should look at the distribution of income after accounting for taxes and transfer payments, not before. At first, this was not a focus of the relatively recent literature on income inequality, although it has come to be much more common. A general conclusion from this additional step is that the US system of taxes and transfer payments lessens the degree of inequality, whether measured by top income shares, Gini coefficients, high-low income ratios, or some other summary measure. But there are disagreements about the extent to which the fiscal system has offset the rise in pretax income inequality in recent decades. For example, Auten and Splinter (2024) find that taxes and transfers have largely offset rising pretax inequality, while Piketty, Saez, and Zucman (2018), who estimate a larger increase in pretax income inequality, conclude that they have not.

While focusing on income after rather than before taxes and transfers is generally a step forward in the analysis of inequality, there are several complications one must confront in making the transition between the two measures.

[1] In what follows, I will refer to "income before taxes and transfers," "before-tax income," and "pretax income" interchangeably, unless otherwise noted.

Incidence Assumptions

To incorporate the effects of government policies on incomes, one needs to make assumptions regarding incidence for both taxes and transfer payments. On the tax side, two prominent examples are the corporate income tax and the payroll tax. Although the influential analysis of Harberger (1962) for many years led to the assumption that the corporate tax was fully borne by owners of all domestic capital, a more common recent approach, reflecting in part the growing importance of international capital flows, has been to assume that the corporate tax leads to a lower level of domestic investment, and in this way some of the burden of the tax is shifted to labor in the form of lower wages. For this reason, both the Congressional Budget Office (2023a) and Auten and Splinter (2024) assign 25 percent of the corporate income tax to labor. For the payroll taxes that fund Social Security and a portion of Medicare, it is common to assume that the full payroll tax, both the share imposed on employees and the share imposed on employers, is borne by labor.

Both assumptions have an impact on estimates of *before-tax* income, given that after-tax income is what we actually observe. For the payroll tax, the measure of before-tax labor income consistent with assumed incidence is gross of all payroll taxes, not just the employee-assessed share. For the corporate income tax, before-tax labor income must include the 25 percent of the corporate tax that labor is assumed to bear.

In this way, tax incidence assumptions do not affect measured after-tax inequality, but because they affect the estimated pretax income distribution, they also affect estimated inequality in the pretax income distribution, and also the change in inequality attributed to taxes; that is, incidence assumptions affect one's conclusion about the progressivity of the overall tax system. There is no neutral approach to the question of incidence that allows one to avoid making assumptions. Of course, it would be convenient to posit that there is no shifting of corporate or payroll taxes, because then the before-tax income distribution corresponds to what is observed, but this convenience comes at the cost of deviating from more empirically grounded assumptions about incidence.

The Valuation of Transfer Payments

When calculating income after taxes and transfers, it is common practice to add transfer payments, whether cash or in-kind, at their cost. Even for cash payments, this step is not as obvious as it may seem. An important issue recognized in the literature involves program take-up—the extent to which individuals who qualify for transfer programs (or tax benefits like the Earned Income Tax Credit) actually opt into the programs. Evidence suggests that the gaps between eligibility and take-up may be significant (Kosar and Moffitt 2017). If stigma or transaction costs limit take-up (Currie 2006 suggests the latter are likely more important), then even those who choose to receive benefits may experience a smaller gain than the cash value of the benefits suggests, after accounting for transaction costs.

For in-kind benefits, the valuation issue is potentially much more significant. Studies of the value that individuals place on the receipt of government health care

insurance, for example, find that the value recipients place on these benefits can be substantially below cost (Finkelstein, Hendren, and Shepard 2019), in part because there is some health care available even to those who lack insurance, so that part of the government benefits ultimately accrue to health care providers. Studies evaluating the distributional effects of government taxes and transfers typically do not adjust for these issues of valuation, but it is worth keeping the issue of who benefits and how much in mind when considering alternative policy actions aimed at addressing inequality. After all, Medicaid spending is by far the largest spending program for the poor, totaling $871.7 billion in 2023.[2] In particular, rising expenditures on government-provided health care benefits might overstate the extent to which these expenditures increase the well-being of recipients, which might suggest a need for further redistribution, but also perhaps a change in the type of redistribution.

Redistribution versus Predistribution

A group of government policies sometimes called "predistribution" has the effect of changing the distribution of income without the use of taxes and benefits. Examples might include government job guarantees, minimum wages, trade protection, or other regulatory interventions. As noted earlier, a standard assumption is that a dollar of wages provides the same benefit as a dollar of transfer payments. This equivalence has long been subject to challenge—for example, through the previously discussed factors limiting the take-up of certain transfer program benefits. However, survey evidence confirms that, especially among less educated voters, policies of predistribution are preferred by individuals to policies that redistribute via the more direct mechanism that does not rely on intervention in markets. This preference may be a possible explanation for the realignment of US political parties among such individuals in response to Democrats' shift from predistribution toward redistribution in their policies (Kuziemko, Longuet-Marx, and Naidu 2023).

Indeed, even though European countries have a lower degree of inequality before taxes and transfers than the United States, this difference is much less after taxes and transfers, because the US tax and transfer system reduces income inequality by more than the comparable systems in Europe (Blanchet, Chancel, and Gethin 2022). This difference suggests a stronger preference in Europe for the predistribution approach.

The borderline between predistribution and redistribution policies is not precise. One would expect the most central element of standard redistribution policy, the progressive income tax, to influence the before-tax income distribution through upper-income taxpayers' responses to potentially high marginal tax rates. This possibility motivates Blanchet, Chancel, and Gethin (2022) to consider the hypothesis that the more unequal US pretax income distribution may be significantly attributable to the decline in top US marginal tax rates, though they find that this is not the case. But there is a key difference between standard redistributive

[2] This figure is according to the Centers for Medicaid and Medicare website at https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet.

policies, such as the progressive income tax, and policies that target the distribution of income before taxes and transfers. While the former policies introduce distortions through explicit and (via means-testing) implicit marginal income tax rates, predistributive policies typically involve additional distortions as well.

For example, import tariffs are economically equivalent to the combination of taxes on the consumption of commodities subject to the tariffs and equal-rate domestic production subsidies for the same commodities, which offset the consumption taxes for domestically produced goods. While there are circumstances in which a subsidy for domestic production might be desirable, such as local learning-by-doing externalities, the standard economic analysis in the absence of such externalities calls for eschewing production distortions, which in this example take the form of favoring domestic production (Diamond and Mirrlees 1971).[3] This is because production distortions reduce the resources available for redistribution without improving the government's options for engaging in redistribution. A similar analysis would apply to the provision of government-guaranteed jobs, which can be interpreted as providing implicit wage subsidies to one sector of the economy. It is generally hard to justify policies aimed at predistribution based on standard economic analysis.

In short, there is possibly a large economic price to be paid for policies aimed at predistribution, in the form of additional economic distortions. The optimal policy, presumably, would need to balance the economic costs of these policies and the potential improvements in well-being experienced by individuals associated with more positive views of income received through market outcomes rather than from the government. One would also have to consider the extent to which such preferences for predistribution are stable.[4]

The Private-Public Borderline

Comparing sources of economic inequality across countries, or within a country over time, can be confounded by differences in the scope of government activity. For example, a country with a public pension system, such as the US Social Security program, would show the receipt of pension payments as government transfers to beneficiaries, while in a country relying on private provision of retirement income through employer-provided pensions, such income would show up in pretax income. Here, too, a focus on income after taxes and transfers would overcome the ambiguity, as in both cases the income would be counted, but, again, the effects of government policy would appear to be quite different, even if the underlying outcomes were the

[3] Note that policies aimed directly at redistribution could also potentially serve to address externalities, for example if higher marginal tax rates among those at the top of the income distribution serve to reduce unproductive rent extraction by corporate executives (Piketty, Saez, and Stantcheva 2014).

[4] Some policies aimed at predistribution might lack even such a trade-off. For example, Autor et al. (2024) find that the tariffs adopted during the first Trump administration failed to generate increases in US employment in the industries targeted, and that the resulting retaliation by foreign governments had negative employment effects on US employment in other industries. The only trade-off present in this case appears to have been with respect to voting outcomes, not actual economic improvements.

same—that is, if both government and private pension systems had the same patterns of payments by individuals and benefits received by individuals.

One approach to overcoming this difference in treatment of otherwise equivalent policies would be to reclassify certain government programs as being, for the purposes of measuring inequality, outside the scope of government activity. In fact, the Congressional Budget Office adopts this approach in its treatment of broad-based transfer programs, including Social Security and Medicare, but this adjustment has its own limitations. Using Social Security as an illustration, this program combines an actuarially fair contributory retirement scheme with a redistribution program, given that replacement rates (or, equivalently, effective rates of return on payroll taxes in terms of subsequent benefits received) are typically found to be higher for low-income individuals (for example, Brown, Coronado, and Fullerton 2009). Treating the entire program as if it were outside the scope of government would lump the redistribution component of Social Security in the pretax distribution of income, rather than attributing it to government policy, thereby understating the magnitude of government policy redistribution.

In summary, among annual measures of income, the distribution of income after taxes and transfers is the most relevant measure of income inequality. But income before taxes and transfers is also relevant, in part because many people do have a preference for what they view as "earned" pretax income over equivalent income from transfer payments. Also, the effect of government policies in redistributing income requires an estimate of income both before and after taxes and transfers, which in turn will depend on assumptions about the incidence of government programs and their valuation on the part of individuals.

Age-Specific Heterogeneity and Life-Cycle Effects

Most analyses of economic inequality compare the income or wealth of individuals or households at a certain time, often a calendar year. However, aggregating individuals of different ages and accounting only for current outcomes can result in very misleading conclusions regarding both the degree of inequality and the extent to which government policy reduces this inequality. This is illustrated below using the example of Social Security. Looking at inequality of income or wealth on an annual basis also leaves no obvious way to combine results regarding wealth inequality and income inequality. One's intuition might suggest that, for a given degree of income inequality, greater wealth inequality signals greater overall economic inequality, and likewise for greater income inequality, conditional on a given degree of wealth inequality. But how should one combine measures of inequality of these two variables, one (wealth) a stock and the other (income) a flow?

Questions about Progressivity

A "snapshot" analysis that looks only at income outcomes at a given year can be misleading about the progressivity of the fiscal system for three reasons. First,

patterns of taxes relative to income are different on an annual basis from such patterns over a longer horizon. A classic illustration of this phenomenon involves the progressivity of consumption taxes. Under the classic permanent-income/life-cycle hypotheses, one would expect greater fluctuations over time in income than in consumption. As a result of consumption-smoothing, annual taxes on consumption would tend to be higher relative to income among those with lower incomes than those with higher incomes (at least to the extent that those income differences are not very persistent). Empirically, this results in consumption-based taxes being less regressive on a lifetime basis than on an annual basis (Poterba 1989).

Second, the age profiles of specific taxes and transfer payments differ markedly from the age profile of income. For example, taxes on labor income, such as payroll taxes, are concentrated at younger ages than taxes on overall income, and the contrast is even larger in comparison to health and pension transfer program benefits. Looking at taxes and transfers at a given age provides only a partial picture of how taxes and transfers vary with the ability to pay.

Third, standard cross-section analyses at a given time will aggregate different age cohorts into a single group. However, individuals of different ages commonly have different patterns of income and different profiles of taxes and transfer payments, and the aggregation can misstate both the degree of pretax inequality and the progressivity of the fiscal system as a whole. A classic illustration comes, again, from the US Social Security system, which is largely a pay-as-you-go system in which taxes paid by working cohorts are spent immediately on benefits paid to largely retired cohorts. The combination of older and younger cohorts can lead to the distorted impression that retired individuals have lower income than those who are working—even if this difference is absent on a lifetime basis. In addition, the pattern of taxes and transfers among these different cohorts makes the Social Security system look extremely progressive, as on a year-to-year basis it is taxing those who are treated as having higher income to make transfer payments to those classified as having lower income.

A related problem is how to treat private retirement income. Auten and Splinter (2024) and Piketty, Saez, and Zucman (2018) include distributions from retirement accounts in measuring the income of retirees, which has the appeal of making such individuals look less "poor." But treating such payments as income is inconsistent with the standard comprehensive definition of income (like the "Haig-Simons" definition discussed in this symposium by Clarke and Kopczuk), which would treat accretions to wealth in such retirement accounts as income only when they occur, rather than when funds are distributed from the accounts. This inconsistency illustrates the challenge of using current income as a measure of well-being.

The Need to Recognize Economic Equivalences

The tax system is rife with provisions that may be economically equivalent to others but have differences in timing that may affect how distributional effects are measured using cross-section data. In some cases, the equivalence involves no

differences in structure or timing at all, but differences in how policies are labeled can still affect how burdens are distributed.[5]

An illustration of the first type of equivalence is the two approaches to tax-favored retirement saving. The traditional approach provides up-front tax deductions for contributions to such retirement accounts as IRAs and 401(k)s, but then taxes withdrawals. In contrast, the alternative "Roth" approach that has grown in significance over the years provides no initial deduction but allows all withdrawals to be tax-free. A well-established result is that, if tax rates are unchanging over time, these two approaches to taxation are equivalent, imposing the same present-value burden on taxpayers. But any assessment based on current tax payments will treat these approaches as very different, with the traditional taxation approach appearing to place more of the tax burden on retirees.

To illustrate the second type of equivalence, involving only labeling, consider the "flat tax" as originally proposed by Hall and Rabushka (1983). The tax has two pieces, a personal wage tax and a business cash-flow tax, which is a tax on the difference between the cash a business takes in and the cash it pays out. These two elements together (except for an exemption amount provided under the wage tax) are equivalent to a value-added tax. To gain some intuition as to why, remember that in a value-added tax, each firm is taxed on its total receipts minus what the firm spent on nonlabor inputs, including investment. In the flat-tax, the amount spent on wages is taxed via the personal wage tax, and the rest of the base for the value-added tax is covered by the business cash-flow tax. However, in a standard income accounting process, the burdens of two separate components of the flat tax would typically be allocated according to capital income and wages, rather than to consumption—the base of the value-added tax, which at a moment in time would result in a quite different pattern of tax burdens across individuals.

While there have been attempts to implement adjustments to distributional analysis aimed at recognizing these types of policy equivalences within a short-horizon approach (notably, US Joint Committee on Taxation 1993), this problem has received far too little attention in the recent literature on inequality.

Analytical Approaches to Lifetime Measures

Some of those who produce cross-section/snapshot analyses have recognized the issues raised for understanding inequality by heterogeneity of ages and a life-cycle perspective and sought methods of trying to deal with it. Some of these methods attempt to make adjustments in the context of an annual time frame. The approach taken by Congressional Budget Office, discussed above, of including Social Security and Medicare benefits in pretax income attacks both elements of the problem—it makes retirees look less poor, and it excludes the Social Security system as an element of government redistribution policy. Piketty, Saez, and Zucman (2018) adopt a related approach, including not only Social Security benefits in

[5] The examples that follow are drawn from the broader discussion of tax equivalences in Auerbach (2019).

pretax income, but also unemployment and disability insurance benefits (but not Medicare benefits), and subtracting payroll taxes, in arriving at their measure of pretax income.

But these adjustments, while aimed at overcoming the limitations of a more basic approach, are only partially successful in solving the underlying problem arising from the aggregation of cohorts and short-horizon analysis. They treat any redistribution within cohorts that actually occurs through such social insurance programs as a reduction in pretax inequality, rather than attributing the effect to government policies. Also, these adjustments have a degree of subjectivity in their scope (as illustrated by the differences in the choices about just what adjustments to make).

A more sweeping conceptual approach is to undertake a forward-looking analysis that groups individuals by age cohorts. Such analyses have been done looking at specific programs, for example, for the Social Security system by Brown, Coronado, and Fullerton (2009) and Congressional Budget Office (2006). In a more comprehensive analysis, Auerbach, Kotlikoff, and Koehler (2023) incorporate all significant US tax and transfer programs at both the state and federal levels, finding that, for middle-aged households, a measure of the fiscal system's progressivity based on a comparison of current income and current taxes net of transfer payments understates lifetime progressivity, in part because of the failure to account for the progressivity of future old-age benefits. In that analysis, lifetime net tax rates—the present value of taxes net of transfer payments divided by lifetime resources (wealth plus the present value of future labor income)—are lower than current-year net tax rates—current-year taxes net of transfers divided by current-year income—especially at the bottom of the resource distribution. This lifetime analysis also offers a method of integrating measures of wealth inequality and income inequality, by looking at the inequality of lifetime resources that incorporate both in a consistent manner.

Yet another approach to overcoming some of the shortcomings of standard distributional analysis is to focus on consumption rather than income, based on the argument that consumption is a more direct measure of well-being, and that, for forward-looking households able to engage in consumption smoothing, current consumption may represent a better measure of lifetime resources than current income. Consumption also provides a more accurate measure of resources among those engaged in informal labor markets, for whom tax records and other measures of income may be inaccurate. Historical estimates of inequality in consumption suggest a much weaker upward trend than the trend in income before or after tax (Meyer and Sullivan 2022). Worth noting, though, is that a focus on annual consumption is not really compatible with the standard evaluation of fiscal progressivity, because there is no analogous annual measure of "consumption before taxes and transfers."

In summary, while the degree of inequality in the United States is considerable, measuring the distribution of resources and the impact of the fiscal system on this distribution on an annual basis for aggregated age cohorts may provide misleading estimates of the level and trend in economic inequality and the role of the fiscal

system in addressing it. The next section of the paper considers the implications for high and possibly rising economic inequality on government policy choices.

Fiscal Policy Implications of High Economic Inequality

Determining the desired size and scope of policies aimed at redistribution involves balancing the benefits of additional redistribution against associated economic distortions. A higher degree of underlying inequality will raise the benefits of redistribution.

The Top Marginal Tax Rate

As an illustration, consider the general formula for determining the welfare-maximizing top marginal income tax rate, as in Diamond and Saez (2011):

$$\tau = \frac{1}{(1 + ae)}$$

where e is the taxable income elasticity among those in the top bracket and a captures the share of income at the top of the income distribution by taking the ratio of the total income for those above the top-bracket income threshold to just their income above that threshold.[6] The intuition for this expression is that a higher value of the elasticity e means more negative income responses to a higher tax rate, lessening the amount of revenue raised and the benefit of a rate increase, while a higher value of the parameter a means that the income subject to the distortion of a higher rate is higher relative to the income actually subject to tax at that higher rate (that is, the income above the threshold). A higher value of either parameter reduces the welfare-maximizing top marginal tax rate. The authors present data from 2005 suggesting that the ratio a is about 1.5 for income level above $300,000. They argue that a midrange estimate of the elasticity e would be 0.25. Those parameter choices imply an optimal income tax for those with the top levels of income of 73 percent. Here, the point to emphasize is that as income inequality increases, represented by a thicker right tail of the income distribution, the parameter a will be smaller—there is more revenue to be gained from taxing very high incomes, improving the trade-off between revenue and the income tax rate's behavioral distortion.

While the basic optimal income tax model envisions governments using tax revenues to engage in redistribution, governments also raise revenues for direct spending on public goods. Within this broader scope of government, a desire for redistribution must compete with other government activities for available

[6] The general formula for the welfare maximizing top marginal tax rate would also include a term measuring the social welfare weight for those in the top bracket. Under a typical social welfare function, this weight is very small and has a negligible effect on the result, which arises when the top welfare weight is zero and is equivalent to the revenue-maximizing top marginal tax rate. Also note that, although it is common to interpret the model as applying to annual income, this has the same shortcomings as analyzing inequality on an annual basis (as discussed in Kaplow 2024).

funds—after all, money spent on redistribution cannot be spent on public goods, and, because of the associated deadweight loss, funds raised through progressive taxation have a higher economic cost than those raised through more efficient tax mechanisms.[7] Ultimately, of course, this tradeoff occurs through the political process, not through optimal tax and spending formulas.

Economic Inequality and the Political Process

Inequality could affect political outcomes in several ways, and the predictions are not always consistent. Perhaps the most straightforward prediction comes from the classic model of Meltzer and Richard (1981), which emphasized the effects of the gap between average income and the income of the median voter. With a progressive income tax, the median voter stands to gain more from an expansion of government the more skewed is the income distribution, and the higher is the ratio of aggregate revenue raised to the tax burden imposed on the median voter. According to this class of models, an increase in inequality should lead to a larger government, with the median voter benefiting at the expense of the rich. By contrast, one concern about inequality (for example, as expressed in Saez and Zucman 2019) is that a high concentration of resources among the few can lead to undemocratic outcomes, as those with high incomes or wealth exert influence over politicians to achieve their desired policy outcomes. Under this view, as inequality has increased, policy decisions have favored the rich, as through the adoption of favorable tax provisions.

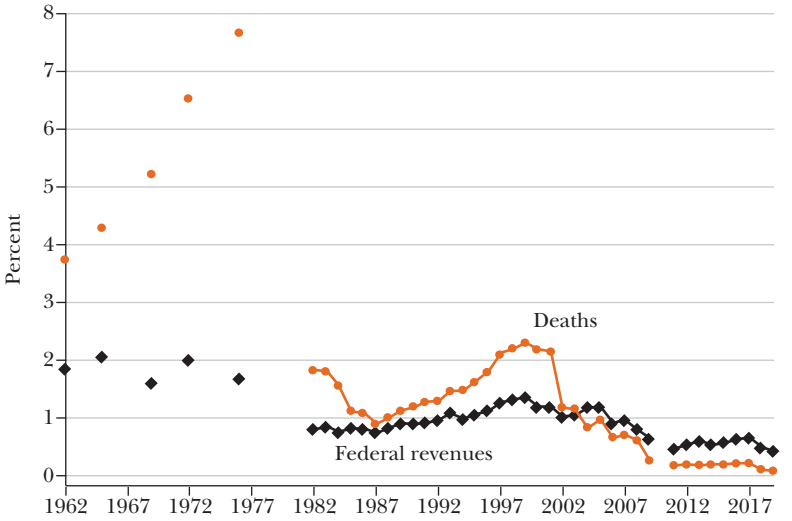
One can find policy outcomes consistent with each of these predictions. The gradual weakening of the US estate tax seems consistent with growing power of the rich. As shown in Figure 1, the estate tax accounted for roughly 2 percent of federal revenues in the 1960s and 1970s, with as many as over 7 percent of decedents subject to the tax.[8] Sharp reductions in the estate tax, introduced through tax legislation in 1981, 2001, 2010, and 2017, have resulted in only 0.08 percent of decedents being subject to the estate tax, which accounted for just 0.4 percent of federal revenue in 2019. It is notable that the rise in the share of decedents subject to the estate tax during the 1960s and 1970s was not associated with an increase in the share of federal revenue accounted for by the estate tax, presumably reflecting the fact that those on the margin of paying the estate tax accounted for little additional tax revenue. Following the same reasoning, the decline since 2001 in the share of federal revenue has been much smaller than that in the share of decedents subject to tax.

On the other hand, over roughly the same period as the estate tax was withering, means-tested transfers (not including benefits provided through the tax

[7] As an extreme illustration of this point, if individuals were identical, with no pretax inequality, there would be no need to devote tax revenues to the funding of redistribution and governments could use uniform nondistortionary lump-sum taxes to finance spending on public goods.

[8] For years before 1982, the IRS provides data for the estate tax only for selected years. It also does not provide estate tax data for 2010 because of the complexity of calculations for that year, when the estate tax was temporarily repealed.

Figure 1
U.S. Estate Tax: Percent of Deaths Subject to Tax and of Federal Revenues



Source: Internal Revenue Service (2024) and Congressional Budget Office (2024).
Note: This figure shows the evolution over time in the share of decedents subject to estate tax and the share of federal revenues accounted for by the estate tax.

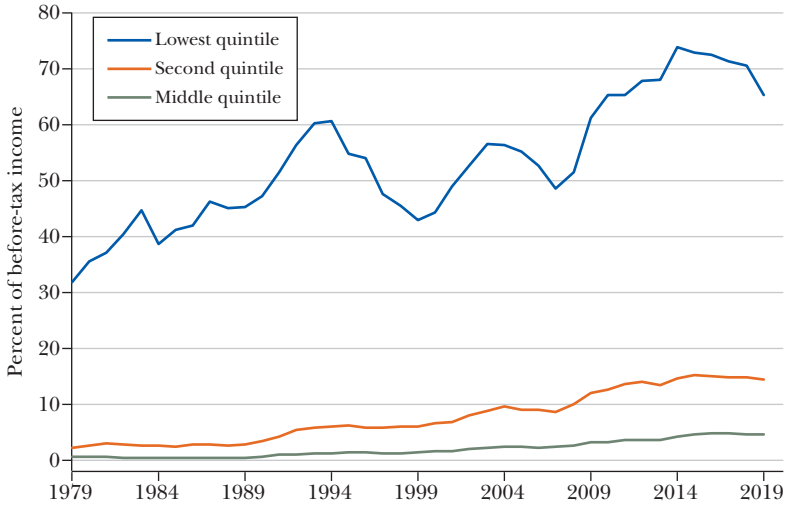
system) were growing steadily relative to before-tax income, as shown in Figure 2, including among those in the middle quintile of the income distribution, which might be relevant for the median voter.

There may be ways to reconcile these fiscal policy developments. For example, the median voter may be generally uninformed about the estate tax, which could be of particular importance to the wealthy; indeed, survey evidence from Kuziemko et al. (2015) suggests that voters' understanding of the estate tax is limited and that their support for expanding it is sensitive to the receipt of information. Also, some of the increase in means-tested benefits reflects the general increase in health care spending. But the contrast in outcomes illustrates the complexity of modeling the impact of inequality on the tax and transfer policies governments adopt.

The Growth of Income-Specific Policies

Adjustments to progressive income tax rates happen once or twice a decade. Sometimes the changes are made over the entire rate schedule; in other cases, the changes are primarily or exclusively at the top. Examples of the former (all tax reductions, in Republican administrations) are the 1981 Economic Recovery Tax Act, the 2001 Economic Growth and Tax Relief Reconciliation Act, and the 2017 Tax Cuts and Jobs Act. Examples of the latter (high-income tax increases, in Democratic administrations) are the 1993 Omnibus Budget Reconciliation Act, which introduced two new higher tax brackets, with the top rate applying to incomes

Figure 2
Average Means-Tested Transfer Rates among Selected Income Groups



Source: Congressional Budget Office (2023b).
Note: This figure shows changes over time in average means-tested transfer payments relative to before-tax income for the bottom three quintiles of the income distribution.

over $250,000, and the 2012 American Taxpayer Relief Act, which permanently extended the 2001 tax rate cuts except for those with taxable incomes over $400,000 ($450,000 for couples).

However, over the past several decades, a period roughly coinciding with the growth in top before-tax income shares, a different approach to increasing progressivity has taken root. This approach, practiced during both Democratic and Republican administrations, has involved increasing tax liabilities at the top through indirect means not involving a change in the stated marginal tax rate, either by phasing out individual tax benefits or by imposing fees or taxes outside the individual income tax. Table 1 summarizes several examples of this approach: some of these provisions raise income taxes and marginal tax rates for those with higher incomes by phasing out income tax exclusions, deductions, and credits (1983, 1990, and 1997); some raise other taxes based on income (2010); and some raise "nontax" social insurance premiums based on income (2003 and 2010). Similarly, although this provision did not become law, the Biden administration proposed in 2024 to increase the Medicare payroll tax and to raise the "net investment income tax" to 5 percent, but only for those above $400,000 in income.

In none of these cases was there an apparent intent to target specific elements of behavior with this choice of an indirect approach. For example, the higher Medicare premiums for high-income individuals were not aimed at reducing their health-care spending, nor were the phased-out child tax credits designed to

Table 1
Indirect High-Income Tax Increases

Year	*Legislation/Proposal*	*Provision(s)*	*Incomes ($) Above (Single/Married)*
1983	Social Security Amendments	Income tax on 50% of Social Security benefits	25,000/32,000
1990	Omnibus Budget Reconciliation Act	Itemized deduction phase-out	100,000
		Personal exemption phase-out	100,000/150,000
1997	Taxpayer Relief Act	Child tax credit with phase out	75,000/110,000
2003	Medicare Modernization Act	Income-based Medicare Part B premiums (IRMAA)	80,000/160,000
2010	Affordable Care Act	0.9% Medicare payroll tax surcharge	200,000/250,000
		3.8% Net investment income tax	200,000/250,000
		Income-based Medicare Part D premiums (IRMAA)	85,000/170,000

Note: IRMAA stands for Income-Related Monthly Adjustment Amount. This table only includes the initial introduction of various provisions into law, not subsequent changes in thresholds, rates, or other related provisions.

discourage fertility among the rich. The provisions in Table 1 are all effectively tax increases on higher income levels, although sometimes with quirky characteristics that can result in seemingly capricious outcomes. As one example, the phase-out of personal exemptions introduced in 1990 increased taxes over the phase-out range in proportion to the number of personal exemptions, thereby imposing larger tax increases on larger families. For cases in which old-age benefits were effectively reduced—the taxation of Social Security benefits and the increase in Medicare premiums—the aim may have been to cut benefits in a progressive manner without being seen as tampering with politically popular programs, although this would have been an effective alternative only if the equivalence was not generally recognized. One might attribute the increase in Medicare taxes to the desire to increase dedicated revenues in a progressive manner, again without appearing to disturb the nature of the program as providing universal social insurance. But what can explain the remaining provisions in Table 1, which approximate income tax increases, but with greater complexity and possibly unintended consequences?[9]

On one hand, the rhetoric associated with the introduction of these indirect provisions has often included arguments that those with higher incomes are able to bear the burden of higher taxes, a rationale fully consistent with a high or increasing share of before-tax income being earned by those at the top. On the other hand,

[9] As an illustration of unintended consequences, the phase-out of itemized deductions, structured so that for almost all taxpayers it was equivalent to an income tax surcharge (because the ceiling on deductions being phased out was not reached), was often misinterpreted by commentators as reducing the value of marginal tax deductions and therefore discouraging deductible activities such as charitable contributions. For discussion, see Viard (2015).

such reasoning can be (and has been during the period covered in Table 1, in 1993 and 2012) used to increase income tax rates at the top directly, rather than indirectly.

Perhaps as with the use of taxes and premiums to reduce social insurance benefits indirectly, the potentially lower salience of indirect increases in tax progressivity has played a role. But for this factor to help explain their observed use, indirect tax increases would need to be more salient among those favoring increases in progressivity than among those opposed. For example, the 1990 high-income tax increases (which also included a 10 percent tax on luxury automobiles, yachts, private-use aircraft, jewelry, and furs) were the result of a deal between President George H. W. Bush and Congress, the deal in which Bush famously reneged on his campaign promise, "Read my lips: no new taxes." It is reasonable to argue that the president leaned toward these indirect means as a way of making the change in course less obvious, at least to those disposed to oppose progressive tax increases. But given the political blowback that Bush experienced, did increasing taxes in this indirect matter prompt less-negative reactions among opponents than a more straightforward approach would have?

In short, increasing inequality may induce a stronger divergence in views of optimal progressivity, with increasing tension between the preferences of the median voter pushing for more progressivity and the political power of the rich rising in opposition. In this setting, differences in relative salience of indirect tax increases could take on greater importance and help explain their growth in recent decades. Alternatively, the growing use of indirect tax increases on the rich may be a political innovation unrelated to inequality or may relate to rising inequality through some other channel—for example, through different views of the permanence of tax increases not based directly on individual income tax rates.

Makers and Takers

During the 2012 presidential campaign, Republican nominee Mitt Romney caused a stir with the release of his critical private comments about the estimated 47 percent of US tax units that did not pay federal income taxes (a number, for 2009, drawn from Williams 2009). While this statistic may have been a surprise to many, it is not particularly puzzling. It results from the combination of relatively low incomes in the bottom half of the income distribution, low marginal tax rates applicable to such income, and the range of important tax benefits available to reduce tax liabilities further for individuals with such incomes, including the standard deduction, personal exemptions, the child tax credit, the earned income tax credit, and the limited taxation of Social Security benefits. Note that the share of individuals not paying taxes would have been lower had various tax benefits, such as the child tax credit, been provided as direct subsidies rather than through the tax system as "tax expenditures."

Mirroring the concern about the large share of the population not paying income taxes is one about the large share of taxes paid by the relatively small population of high-income individuals. One recent estimate finds that taxpayers in the

top income percentile accounted for 46 percent of all federal income tax payments in 2021 (Tax Foundation 2024). Of course, the share of income taxes paid by high-income taxpayers is a function of both the tax structure and the group's share of overall income. Even for a given tax structure, increasing inequality at the top also increases the share of taxes paid by those at the top.

In one sense, concern about the implications of the high share of households not paying income taxes and the high share of income taxes paid by those at the top is simply a recasting of the above discussion about the effects on the political equilibrium of the widening gap between median and average incomes. From this perspective, the fact that—as many critics of the focus exclusively on income taxes have pointed out—a large share of low-income households pays other taxes, notably payroll taxes, may not be especially relevant if one is considering decisions about increasing the income tax to fund government expansion. At least some who are unhappy with the current situation have expressed concern about it promoting a lack of voter engagement, leading to proposals to require all households to pay at least some income tax, even if only a token amount.[10]

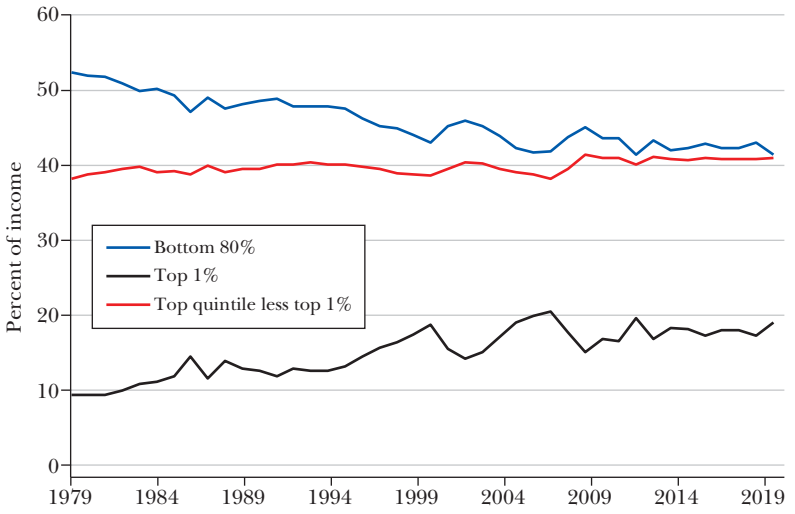
Revenue Volatility

Pretax income is more volatile at the top of the income distribution. Figure 3 illustrates this point for households sorted according to their "market income," which is the Congressional Budget Office measure of income before all taxes and transfers. The figure displays income shares for three groups between 1979 and 2020, based on cross sections of income in the respective years: the bottom four quintiles, the top quintile excluding the top percentile, and the top percentile. Not only is the top percentile's income share the most volatile, but its pattern over time is different, peaking in booms and falling in recessions, notably those in 1990 and especially 2001 and 2008–2009 (although not those in the early 1980s). This greater volatility at the top reflects the fact that the sources of income at the top, capital income (including capital gains and business profits), executive compensation, and self-employment income, are more volatile than wage and salary income. Even though this result holds for just the top income percentile, this group accounts for a significant share of federal income taxes, as already discussed.

Thus, assuming that the cyclical behavior of the income of those at the top does not change as their share of income grows, increasing inequality would lead to stronger procyclical volatility of income tax revenues, holding the tax structure constant. This effect is potentially large. For each dollar of aggregate income fluctuation between 1979 and 2020, around one-half has been accounted for by the income of those in the top percentile, so increased income subject to the same swings could

[10] Such a proposal, which generated heated reaction, was put forward in 2022 by Senator Rick Scott (R-Florida), with the language: "All Americans should pay some income tax to have skin in the game, even if a small amount." See https://rescueamerica.com/wp-content/uploads/2022/02/RickScott-11-Point-Policy-Book.pdf.

Figure 3
Shares of Market Income by Market Income Quantile



Source: Congressional Budget Office (2023c).
Note: This figure shows the shares of market income over time for the bottom four quintiles of market income, the top quintile of market income excluding the top 1 percent, and the top 1 percent.

substantially increase aggregate income volatility.[11] Further, if increasing inequality brings with it an increase in the progressivity of the income tax schedule and higher marginal tax rates at the top, this change would further exacerbate the increase in the volatility of income tax revenues.

While other factors in the past have contributed to greater revenue stability, notably the decline in top marginal tax rates in the 1980s (Kniesner and Ziliak 2002) and the shift in federal tax revenue composition from corporate income taxes to payroll taxes, these factors are unlikely to play an important role in the future, given the growing share of nonworking adults due to population aging, the recent relative stability of corporate income taxes as a share of federal revenues, and the political pressure toward progressive income tax increases targeting high-income individuals.

Revenue volatility and its potential increase is an issue at the state level as well, particularly in states that rely heavily on a progressive income tax and have income inequality that mirrors (or exceeds) the national level. For example, the state of California, with a very progressive state income tax (including full taxation of capital gains) and a relatively unequal distribution of income, has experienced years of large deficits and large surpluses in rapid sequence (Auerbach 2010). For states,

[11] This estimate is based on a simple methodology discussed in the online Appendix.

such fluctuations can be very challenging, due to balanced-budget requirements and the lack of adequate "rainy day" funds to cushion annual fluctuations in budget gaps. Although the lack of annual budget limits makes the problem less serious at the federal level, it is still a possible cause for concern, particularly given the potential for government decisions to focus on the short run.

Higher Inequality and Higher Tax Rates at the Top

Given high inequality and the aim of imposing a progressive tax burden, the government faces the prospect of trying to raise a substantial share of federal revenues from those at the top of the income distribution, using potentially high marginal tax rates. Evidence suggests that such individuals are more sensitive to tax rates and have greater access to tax avoidance strategies (Auerbach and Siegel 2000; Gruber and Saez 2002), so simply increasing marginal tax rates raises less revenue than a simple calculation would suggest. Here, there are two broad options: reform of the existing tax system or adopting alternative approaches to raise revenues in a progressive manner.

Perhaps the clearest target for reform of the existing system is the treatment of capital gains, which are highly concentrated at the top of the income distribution and, by being taxed only upon realization and not at all at death, provide ample opportunities for tax avoidance. Some have suggested that raising capital gains tax rates would generate substantial additional tax revenue (for example, Sarin et al. 2022), although that conclusion remains controversial (Dowd and Richards 2021). But other reforms of the capital gains tax could sharply reduce the scope for taxpayer avoidance responses and be more effective at raising tax revenue.

Two obvious reforms would be a change in the tax treatment of gains at death (either taxing them at death or collecting tax when the assets are sold by those who inherit them) and moving toward taxing capital gains on accrual rather than realization. While there are challenges to taxing capital gains on accrual for illiquid and hard-to-value assets, methods to modify realization-based taxation to accomplish similar objectives—in particular, removing the incentive to defer realization through an effective interest charge on deferred taxes—have been developed (for example, Auerbach 1991).[12] A key issue regarding how much revenue could be collected through a move to accrual or accrual-equivalent taxation of capital gains is how gains accumulated prior to enactment would be treated. Taxing such gains immediately or over the very short run would amount to a large, one-time, unannounced wealth tax, which could be attractive because of its apparently lump-sum nature but could also influence expectations about future tax policy and hence behavior. A

[12] Such "retrospective" taxation might also be preferred to accrual taxation on constitutional grounds. Even deeper constitutional concerns have been put forward concerning a federal wealth tax, which some have seen as an alternative to increasing capital gains taxes (for discussion, see Hemel 2019).

scheme of this form was proposed for very high-wealth individuals during the Biden administration (Saez, Yagan, and Zucman 2021).

An alternative approach could involve a shift to consumption taxation. While consumption taxation in the form of a value-added tax is a tool used in virtually all major economies, it has never been adopted in the United States. One objection to US adoption has been the perceived regressivity of a value-added tax, particularly if adopted broadly and covering necessities such as food, clothing, and shelter. This has not stopped the widespread use of value-added taxation in other countries, based on the rationale that additional tax revenues may then be used to fund progressive social safety net programs. Even so, if a consumption tax is instead implemented in the form of a personal expenditure tax with a progressive rate structure, rather than in the form of a value-added tax, there need be no concern at all with regressivity. This is particularly so if the tax is implemented as a supplement to the existing tax system and applied only to high-income individuals or households, as proposed by Andrews (1980) and others.

A major benefit of imposing a supplemental tax on expenditures of high-income individuals is that, because it is based on cash flows (income less net saving), it is not subject to the difficulties of measuring and taxing capital income that play a major part in tax avoidance through such activities as the deferral of capital gains and the operation of closely held businesses. Depending on transition provisions, adoption of an expenditure tax could also impose a tax on the consumption financed by previously accumulated wealth, thereby effectively imposing a one-time tax on this wealth in much the same way that a tax on previously accumulated capital gains would. An additional benefit of reliance on expenditure taxation rather than other methods of taxing high-income individuals would be the reduction in revenue volatility, as consumption fluctuates less than income from year to year, particularly when income includes the volatile component of realized capital gains. Auerbach (2009) discusses the advantages of consumption-based taxation in greater detail.

Like a personal expenditure tax, a shift in the corporate income tax in the direction of taxation based on cash flows and the location of consumption could reduce the opportunities for cross-border tax avoidance among multinational companies, which has been seen as a major problem standing in the way of progressive taxation (Auerbach 2017).

Finally, as discussed earlier in relation to the taxation of Social Security benefits and the use of income-based Medicare premiums, modifications of the tax system and benefit-program premiums can serve as a substitute for more direct changes in benefit schedules that might be desirable in light of substantial economic inequality, including among the elderly. This approach may offer dedicated revenues needed to maintain program viability. Further adjustments in this direction are certainly possible, for example, through fuller taxation of Social Security benefits or more progressive premiums for Medicare Parts B and D. At some point, though, such changes could increase progressivity to the point of altering what has been deemed by program defenders to be a politically useful perception—if already rather inaccurate—of Social Security and Medicare as "universal" benefit programs,

earned through years of labor force participation. That is, there may be limits to the extent to which these programs can engage in redistribution without coming to be perceived as primarily low-income transfer programs.

There is substantial economic inequality in the United States. Grappling with how much inequality actually exists, what is being done to reduce it, and what additional steps might be done requires digging into the practical details of how taxes and benefits are designed. Of course, the prospects for such reforms must be assessed not only with respect to their technical feasibility, but also within a fraught political setting to which inequality has undoubtedly contributed.

■ *I am grateful to the editors, Timothy Taylor, Jonathan Parker, Nina Pavcnik, and Heidi Williams, and to Jerry Auten, Jim Hines, Louis Kaplow, Wojciech Kopczuk, Emmanuel Saez, Joel Slemrod, David Splinter, Danny Yagan, and Gabriel Zucman for very helpful comments on an earlier draft.*

References

Andrews, William D. 1980. "A Supplemental Personal Expenditure Tax." In *What Should Be Taxed, Income or Expenditure?*, edited by Joseph A. Pechman, 127–51. Brookings Institution.

Auerbach, Alan J. 1991. "Retrospective Capital Gains Taxation." *American Economic Review* 81 (1): 167–78.

Auerbach, Alan J. 2009. "The Choice between Income and Consumption Taxes: A Primer." In *Institutional Foundations of Public Finance: Economic and Legal Perspectives*, edited by Alan J. Auerbach and Daniel N. Shaviro, 13–46. Harvard University Press.

Auerbach, Alan J. 2010. "California's Future Tax System." *California Journal of Politics and Policy* 2 (3).

Auerbach, Alan J. 2017. "Demystifying the Destination-Based Cash-Flow Tax." *Brookings Papers on Economic Activity* 48 (2): 409–32.

Auerbach, Alan J. 2019. "Tax Equivalences and Their Implications." *Tax Policy and the Economy* 33: 81–107.

Auerbach, Alan J. 2025. *Data and Code for: "Public Finance Implications of Economic Inequality."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E218906V1.

Auerbach, Alan J., Laurence J. Kotlikoff, and Darryl Koehler. 2023. "US Inequality and Fiscal Progressivity: An Intragenerational Accounting." *Journal of Political Economy* 131 (5): 1249–93.

Auerbach, Alan J., and Jonathan M. Siegel. 2000. "Capital-Gains Realizations of the Rich and Sophisticated." *American Economic Review* 90 (2): 276–82.

Auten, Gerald, and David Splinter. 2024. "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends." *Journal of Political Economy* 132 (7): 2179–227.

Autor, David, Anne Beck, David Dorn, and Gordon H. Hanson. 2024. "Help for the Heartland? The Employment and Electoral Effects of the Trump Tariffs in the United States." NBER Working Paper 32082.

Blanchet, Thomas, Lucas Chancel, and Amory Gethin. 2022. "Why Is Europe More Equal than the United States?" *American Economic Journal: Applied Economics* 14 (4): 480–518.

Brown, Jeffrey R., Julia Lynn Coronado, and Don Fullerton. 2009. "Is Social Security Part of the Social Safety Net?" *Tax Policy and the Economy* 23 (1): 37–72.

Congressional Budget Office. 2006. "Is Social Security Progressive?" Economic and Budget Issue Brief, December 15. https://www.cbo.gov/sites/default/files/109th-congress-2005-2006/reports/12-15-progressivity-ss.pdf.

Congressional Budget Office. 2023a. *The Distribution of Household Income in 2020.* Congressional Budget Office.

Congressional Budget Office. 2023b. *Data for: Trends in the Distribution of Household Income from 1979 to 2020.* Congressional Budget Office. https://www.cbo.gov/system/files/2023-11/59510-data.xlsx (accessed April 7, 2024).

Congressional Budget Office. 2023c. *The Distribution of Household Income in 2020—Additional Data for Researchers.* https://www.cbo.gov/system/files/2023-11/59509-additional-data-for-researchers.zip (accessed April 9, 2024).

Congressional Budget Office. 2024. "Historical Budget Data, Feb. 2024." https://www.cbo.gov/data/budget-economic-data#2 (accessed April 7, 2024).

Currie, Janet. 2006. "The Take-Up of Social Benefits." In *Public Policy and the Income Distribution*, edited by Alan J. Auerbach, David Card, and John M. Quigley, 80–148. Russell Sage Foundation.

Diamond, Peter A., and James A. Mirrlees. 1971. "Optimal Taxation and Public Production I: Production Efficiency." *American Economic Review* 61 (1): 8–27.

Diamond, Peter, and Emmanuel Saez. 2011. "The Case for a Progressive Tax: From Basic Research to Policy Recommendations." *Journal of Economic Perspectives* 25 (4): 165–90.

Dowd, Tim, and Zach Richards. 2021. "Contextualizing Elasticities for Policymaking: Capital Gains and Revenue-Maximizing Tax Rates." Preprint, SSRN. https://dx.doi.org/10.2139/ssrn.3767121.

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

Gruber, Jon, and Emmanuel Saez. 2002. "The Elasticity of Taxable Income: Evidence and Implications." *Journal of Public Economics* 84 (1): 1–32.

Hall, Robert E., and Alvin Rabushka. 1983. *Low Tax, Simple Tax, Flat Tax.* McGraw-Hill.

Harberger, Arnold C. 1962. "The Incidence of the Corporation Income Tax." *Journal of Political Economy* 70 (3): 215–40.

Hemel, Daniel. 2019. "Taxing Wealth in an Uncertain World." *National Tax Journal* 72 (4): 755–76.

Internal Revenue Service. 2024. "Taxable Estate Tax Returns as a Percentage of Adult Deaths, Selected Years of Death, 1934–2019." https://www.irs.gov/statistics/soi-tax-stats-historical-table-17 (accessed April 7, 2024).

Kaplow, Louis. 2024. "Optimal Income Taxation." *Journal of Economic Literature* 62 (2): 637–738.

Kosar, Gizem, and Robert A. Moffitt. 2017. "Trends in Cumulative Marginal Tax Rates Facing Low-Income Families, 1997–2007." *Tax Policy and the Economy* 31 (1): 43–70.

Kniesner, Thomas J., and James P. Ziliak. 2002. "Tax Reform and Automatic Stabilization." *American Economic Review* 92 (3): 590–612.

Kuziemko, Ilyana, Nicolas Longuet-Marx, and Suresh Naidu. 2023. "'Compensate the Losers?' Economic Policy and Partisan Realignment in the US." NBER Working Paper 31794.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva. 2015. "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review* 105 (4): 1478–1508.

Meltzer, Allan H., and Scott F. Richard. 1981. "A Rational Theory of the Size of Government." *Journal of Political Economy* 89 (5): 914–27.

Meyer, Bruce D., and James X. Sullivan. 2022. "Consumption and Income Inequality in the US since the 1960s." NBER Working Paper 23655.

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

Piketty, Thomas, Emmanuel Saez, and Stefanie Stantcheva. 2014. "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities." *American Economic Journal: Economic Policy* 6 (1): 230–71.

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

Poterba, James M. 1989. "Lifetime Incidence and the Distributional Burden of Excise Taxes." *American Economic Review* 79 (2): 325–30.

Saez, Emmanuel, Danny Yagan, and Gabriel Zucman. 2021. "Capital Gains Withholding." Unpublished.

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

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

Sarin, Natasha, Lawrence Summers, Owen Zidar, and Eric Zwick. 2022. "Rethinking How We Score Capital Gains Tax Reform." *Tax Policy and the Economy* 36: 1–33.

Smith, Matthew, Owen Zidar, and Eric Zwick. 2023. "Top Wealth in America: New Estimates under Heterogeneous Returns." *Quarterly Journal of Economics* 138 (1): 515–73.

Tax Foundation. 2024. *Summary of the Latest Federal Income Tax Data, 2024 Update*. March 13. https://taxfoundation.org/data/all/federal/latest-federal-income-tax-data-2024/.

US Joint Committee on Taxation. 1993. *Methodology and Issues in Measuring Changes in the Distribution of Tax Burdens*. JCS-7–93.

Viard, Alan D. 2015. "The Basic Economics of Pease and PEP." *Tax Notes* 146 (6): 805–10.

Williams, Roberton. 2009. "Who Pays No Income Tax?" *Tax Notes* 123: 1583.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 171–194

A Hitchhiker's Guide to Federal Reserve Participation in Fixed Income Markets

Nina Boyarchenko and Or Shachar

Debt securities are a fundamental asset used both to finance governments and corporations and also as an asset class held by insurance companies, pension funds, and mutual funds. In the United States, most debt market securities—or, more simply, bonds—are issued with a fixed interest rate, leading to debt markets being colloquially called "fixed income" markets. Beyond financing US federal, local, and state governments and US firms, US fixed income markets play a central role in global financial markets. Highly rated US fixed income securities are frequently used as a savings vehicle and as collateral in financial transactions.

Despite the central role that US fixed income markets play in the global economy, the last two decades have also demonstrated that these markets may be subject to systemic stress. What are the features of fixed income markets—even those for the safest securities—that may lead to market distress? What policies are available to the US central bank to support the functioning of such markets during normal times, and what interventions may be used to stabilize fixed income markets during periods of stress?

In this article, we survey the relationship between US fixed income markets and the public sector—in particular, the Federal Reserve—through the lens of the role of key intermediaries common to US fixed income markets, so-called "primary dealers." We begin with an overview of the role of primary dealers in US fixed

■ *Nina Boyarchenko and Or Shachar are both Financial Research Advisors in Capital Markets, Federal Reserve Bank of New York, New York City, New York. Boyarchenko is also a Research Fellow, Center for Economic Policy and Research, London, United Kingdom, and Fellow, CESIfo Research Network, Munich, Germany. Their email addresses are nina.boyarchenko@ny.frb.org and or.shachar@ny.frb.org.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241436.

income markets. We then proceed to discuss the relationship between US fixed income markets, intermediaries in those markets, and how liquidity interventions accomplish different goals than monetary policy. Finally, we discuss the changing structure of the US fixed income markets and argue that, from the perspective of prudential regulators, more tools for quantitative monitoring of market functioning are necessary for early detection of potential fragilities. The other papers in this symposium delve deeper into the particular aspects of US Treasury, US corporate bond, and US municipal bond markets.

Primary Dealers and Fixed Income Markets

US fixed income markets are among the largest in the world, with $28.3 trillion in US Treasuries, $11.2 trillion in US corporate bonds, $4.2 trillion in municipal bonds, and $2 trillion in agency debt outstanding at the end of 2024. They are instrumental for the functioning of the real economy by funding the US government, municipalities, and firms, and for improving the efficiency of the mortgage markets for households. Fixed income instruments also play a crucial role in investors' portfolios by providing reliable source of income, capital preservation, and diversification, among other benefits. Unlike equity securities, holders of debt securities are not owners of the issuing entity, but rather creditors of the issuer, and are entitled in case of the bond's default to claims on the issuer's assets and cash flows, with the priority of their claim determined by the bond's terms.

US Treasuries are issued by the US government and include bills, notes, and bonds. Shorter-term Treasury bills mature in up to 52 weeks and do not make coupon payments. Rather, they are sold for less than their face value but pay their full face value at maturity. The interest earned is the difference between the purchase price and the par value at maturity. For example, if an investor buys a 26-week Treasury bill with a face value of $1,000 for $980, the investor will receive the full $1,000 at maturity. The $20 difference between the purchase price and the face value represents the interest earned on that investment. Treasury notes are issued with maturities of two, three, five, seven, or ten years and pay interest every six months. While two- through seven-year notes are issued each month as new issues, the ten-year note is issued as a new security on a quarterly basis (February, May, August, and November) and typically reopens in the other eight months. Reopenings are additional amounts of previously issued securities with the same maturity date and interest rate. Treasury bonds are issued quarterly with 20- and 30-year maturities (and reopen in the other months) and pay interest every six months.[1]

Agency mortgage-backed securities are financial securities whose payoffs are based on the pool of mortgages "bundled" into the security. US agency mortgage-backed securities are issued by government-sponsored enterprises, and in

[1] The 20- and 30-year Treasury bonds are issued quarterly (February, May, August, and November) and reopen on the other eight months of the year.

case of an individual mortgage borrower default, the securities are guaranteed by the issuing agency—not the full faith and credit of the US government. Because they get implicit support from the US government, they are considered to be of high credit quality. Issuers of agency bonds include the Federal National Mortgage Association (Fannie Mae) and Federal Home Loan Mortgage Corporation (Freddie Mac).

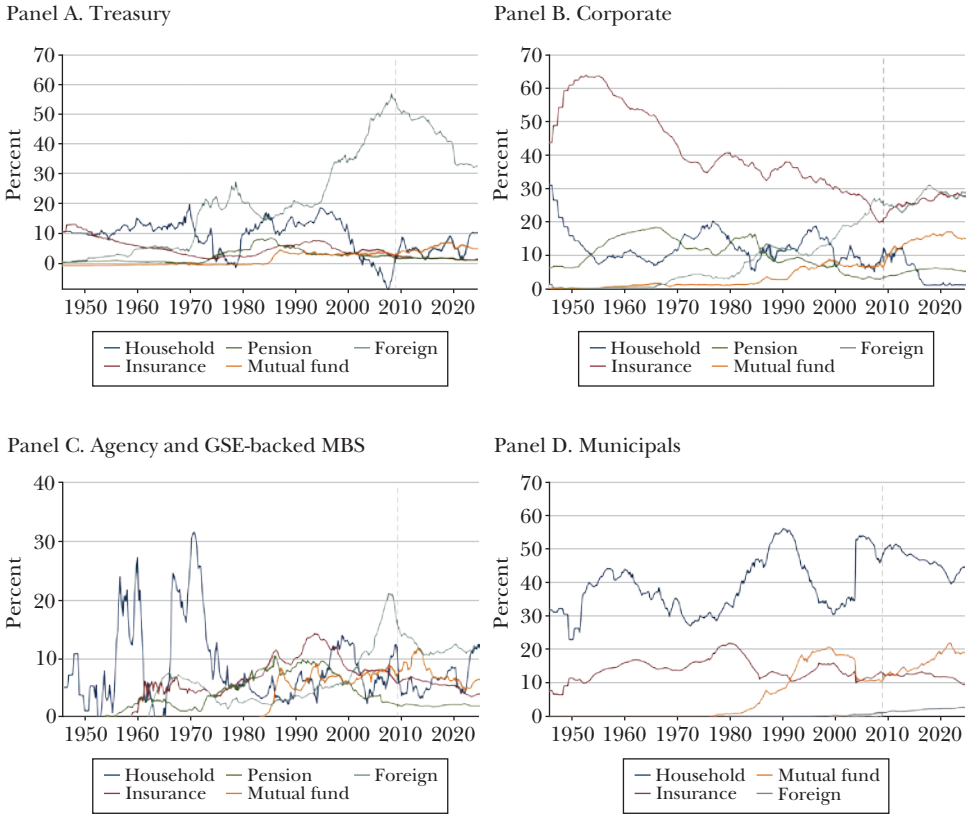
Municipal bonds are debt securities issued by state and local governments to fund public projects or operations, offering investors regular interest payments and the return of principal at maturity, often with tax-exempt interest income. There are two types of municipal bonds: general obligation bonds backed by the "full faith and credit" of the issuer and thus backed by the issuer's tax incomes; and revenue bonds backed by the revenue generated by a specific project, such as a toll road or a utility system.

US corporate bonds are issued by companies to raise capital, typically through an investment bank that underwrites and markets the bond offering to investors, with bonds representing a debt obligation where the company promises to repay the principal and interest at specified times. The market often separates investment-grade bonds and high-yield bonds. "Investment-grade" corporate bonds are issued by companies with credit ratings of Baa3 or BBB or above by Moody's or S&P, respectively, while "high-yield corporate bonds" (also called "junk bonds") are issued by companies with credit ratings of Ba1 or BB+ or below by Moody's and S&P, respectively. In terms of risk, high-yield bonds carry a higher risk of default than investment grade bonds, and therefore offer a higher yield.

Figure 1 shows the fraction of major holders for the four broad fixed income categories on which we focus. Each panel breaks down the composition of holders over time for these asset classes, highlighting how investor participation has evolved. It shows that the major holders are different across the four fixed income categories, and change over time. The vertical dashed line in each panel marks March 2009, corresponding to the S&P 500's low point during the global financial crisis. The crisis led to notable shifts in investor holdings, as seen in the changes following this period.

Panel A focuses on US Treasuries. Since the 1970s, foreign investors have played an increasingly dominant role, with their share rising sharply in the late 1990s and early 2000s. Their holdings peaked around 2009 before gradually declining to early 2000s levels by the end of 2024. Other significant holders include pension funds and mutual funds, though their shares have remained relatively stable over time. Panel B examines corporate bonds, where insurance companies have historically been the largest holders. However, their share has steadily declined, while mutual funds and foreign investors have gained prominence, reflecting broader shifts in investment preferences. Panel C presents the major holders of mortgage securities backed by government-sponsored enterprises. Household holdings were more significant in the earlier decades, but over time, foreign investors, mutual funds, and pension funds have increased their presence in this market. Panel D shows the holders of municipal bonds. Households have consistently been the dominant investors, although mutual funds have gained market share over time. Unsurprisingly, unlike the other securities categories, foreign investors play a minimal role

Figure 1
Fixed Income Major Holders (as a Fraction)



Source: Data are from Financial Accounts of the US—Z.1 (https://www.federalreserve.gov/releases/z1/).
Note: This figure shows the fraction of major holders of fixed income securities, based the flow of funds reports. Panels A, B, C, and D break down Treasury, Corporate, mortgage securities backed by government-sponsored enterprises and municipals holdings, respectively. The dashed vertical line marks March 2009, when the S&P 500 bottomed towards the "end" of the global financial crisis of 2007–2009.

in this market. This is largely due to the tax advantages that municipal bonds offer to US investors, as the interest income is generally exempt from federal taxes, and in many cases, state and local taxes. Because foreign investors do not benefit from these tax exemptions, they have less incentive to hold municipal bonds compared to domestic investors. Taken together, the figure highlights the changing landscape of fixed income ownership, emphasizing how different investor groups have adjusted to economic trends and market events over time.

In the following two sections, we explore how the Federal Reserve implements policy through the prism of the US fixed income markets, highlighting conventional management of the federal funds interest rate, unconventional monetary tools of quantitative easing, and liquidity interventions during times of financial

stress. Here, we emphasize how the shifting landscape of who holds fixed income securities directly affects how effectively conventional and unconventional monetary policies pass-through to financial markets and the broader economy. For example, Federal Reserve's large-scale asset purchases of both Treasuries and agency mortgage-backed securities fundamentally alter the dynamics of demand factors in these markets, potentially changing the way that yields are determined.

More generally, to understand how investors become holders of specific fixed income securities, we need to understand how securities are issued—sold in the primary market—and how they are subsequently re-traded in the secondary market among investors. We start by describing how the US Treasury market operates, as it serves as the reference point for pricing, liquidity, and risk in all other fixed income markets. More importantly, the group of dealers that plays a crucial role in the functioning of both the primary and secondary Treasury markets is also instrumental in the functioning of the broader fixed income markets. The US Treasury securities are offered for sale by the Treasury in the primary market through auctions where all successful bidders pay the same price. The price that everyone pays is the highest bid that still allows all securities to be sold. Treasury auctions are open to all investors, but a select group of financial institutions—"primary dealers"—play an instrumental role in these auctions. Primary dealers have an obligation "to bid on a pro-rata basis in all Treasury auctions at reasonably competitive prices," ensuring that the entire debt issue is sold at a reasonable price.[2] Other investors, such as pensions, insurance companies, mutual funds, and foreign central banks, participate in the auctions either directly or indirectly by submitting bids through a primary dealer.

The origins of the current primary dealer system can be traced back to the late 1950s and was designed "to assure successful financing of the Government's requirements and, at the same time, to minimize monetization of the public debt" (FRBNY Circular no 3665, March 5, 1951). At the end of the 1950s, there were 17 primary dealers who provided the Treasury with a reliable distribution network for its debt securities (Garbade 2016); today there are 24 primary dealers.

In parallel, the auction system for Treasury notes and bonds began taking shape in the 1970s,[3] evolving to complement the primary dealer system. Prior to 1970s, Treasury securities were sold through fixed-price subscription offerings and exchange offerings (Garbade 2004). In a fixed-price subscription offering, the Treasury determined the maturity date and coupon rate of a new issue, specified the total amount available for sale, and invited public subscriptions at a predetermined price. If the price was set too high, the issue might not sell in full, leaving the Treasury with unsold securities and the need to find alternative buyers or adjust future offerings.

[2] This is quoting the "Primary Dealers" page at the website of the Federal Reserve Bank of New York, see https://www.newyorkfed.org/markets/primarydealers. The page also includes a link to the list of all current primary dealers.

[3] Bills have been auctioned in a similar fashion as they are today throughout the 1950s and 1960s (Garbade 2004).

In an exchange offering, the Treasury specified the maturity dates and coupon rates for one or more new notes and/or bonds and invited the public to swap maturing securities for an equivalent principal amount of the newly issued ones. Similar to subscription offerings, setting the price too high could prompt investors to redeem a larger-than-expected portion of the maturing debt, potentially creating cash-flow challenges for the Treasury.

To address the Treasury's challenge in accurately gauging investors' demand and setting a price that would ensure that investors would buy the full amount offered without excessive oversubscription, as well as to enhance market efficiency, transparency, and fairness in the sale of Treasury securities, competitive bidding was introduced in the 1970s. Under this system, institutional investors would submit bids for Treasury securities, specifying the amount they wished to purchase and the interest rate they were willing to accept. It allowed primary dealers to bid for securities in a structured auction format, aligning the interests of primary dealers with those of the broader market. The efficiency of discriminatory price auction format, where winning bidders would pay the price they bid, was called into question when it was found out that Salomon Brothers accumulated 94 percent of the two-year Treasury notes issued in the May 1991 auction. This was a violation of the Treasury regulation that no bidder may bid for more than 35 percent of the issues in any single auction, and led to significantly higher prices of the two-year notes issued in May 1991 than the estimated competitive prices in the four-week post issue period (Jegadeesh 1993).

Following the improprieties at Salomon Brothers, the US Treasury began experimenting with uniform-price auctions in September 1992 for the two- and five-year Treasury notes. The method was later extended to all maturities in 1998 (Malvey and Archibald 1998). In a uniform price auction, bidders submit their offers specifying the quantity of bonds they wish to purchase and the price they are willing to pay. After collecting all bids, the price is determined by choosing the highest price at which all the offered bonds can be sold. All winning bidders pay the clearing price, even if their bids were higher. This encourages bidders to bid their true value, because they know they will not pay more than the uniform price. The auctions are open to primary dealers, their customers (who are effectively indirect bidders), and direct bidders, a group that includes insurance firms, pension funds, foreign central banks, and so on. Primary dealers consistently bid higher yields in the auctions compared to direct and indirect bidders (Hortaçsu, Kastl, and Zhang 2018). The relationship between primary dealers and the secondary market is a vital component of the Treasury market's architecture.

The secondary market for Treasury securities is over-the-counter; that is, these securities do not trade on an exchange.[4] Instead, primary dealers serve as

[4] Treasury bonds were traded in the New York Stock Exchange during the late 1910s and early 1920s. However, in the early 1920s, trading gradually migrated to an over-the-counter market. This transition was partly due to the preferences of institutional investors who found the over-the-counter market a more efficient way to manage and transfer risk (Potter 2015).

intermediaries: they buy many securities at auctions, then resell them to other market participants over time and may choose to hold the remainder of the bonds from issuance through maturity and/or redemptions. They are compensated for taking such inventory risk through subsequent price appreciation of securities bought at auction (Fleming, Nguyen, and Rosenberg 2024). Although the time and the amounts of Treasury auctions are anticipated, Treasury security prices in the secondary market tend to dip in the few days leading up to Treasury auctions and recover shortly thereafter (Lou, Yan, and Zhang 2013). Also, the price impact in the secondary market is more pronounced when primary dealers' risk-bearing capacity is limited: when the auction size is larger, when dealers are more capital-constrained, or when interest rates are more volatile (Lou, Yan, and Zhang 2013).

The expansion of auction participants combined with technological advancements has led to major changes in the secondary market for Treasury securities, especially in the interdealer market segment. The majority of interdealer trading takes place not between dealers directly, but instead through separate firms that run interdealer brokerage platforms using an anonymized and centralized limit order book (Brain et al. 2018). Only primary dealers had direct access to the interdealer brokerage platforms until 1992. However, the interdealer brokers (under pressure from increased competition driving down their commission rates) decided to expand access beyond primary dealers. Starting in 1986, the Government Securities Clearing Corporation (GSCC) used to be responsible for electronic clearing and netting of trades for government and agency debt securities, including both new issues and the sale of existing government securities. It merged with the Mortgage-Backed Securities Clearing Corporation to form the Fixed Income Clearing Corporation in 2003. Thus, when the interdealer brokers decided to expand access to their platforms in 1992, they included firms that were not primary dealers, but which were so-called "netting members" of the GSCC (Fleming, Nguyen, and Rosenberg 2024). More recently, interdealer brokers have granted access to their trading platforms to an even wider range of participants, including participants outside what is now the Fixed Income Clearing Corporation netting membership.

As mentioned earlier, the importance of primary dealers as intermediaries extends to other fixed income markets, such as corporate bonds, municipal bonds, and agency mortgage-backed securities, not only because Treasuries are used as collateral and financing in these markets, but also because of the primary dealers' size, capital commitments, and already established trading networks. Corporate bonds, municipal bonds, and agency mortgage-backed securities are brought to the market by underwriters, who commit to buying the newly issued securities, and at least for a time, to making markets in them. Many of the dominant underwriters are also primary dealers, which underscores their essential role in the broader fixed income ecosystem.

Specifically, agency mortgage-backed securities, which are typically backed by government-sponsored enterprises such as Fannie Mae, Freddie Mac, and Ginnie Mae, are generally sold through negotiated offerings. In these offerings, the government-sponsored enterprises work with underwriters to sell the securities to

investors. The issuance process for agency mortgage-backed securities involves the pooling of residential mortgages and the creation of securities that represent a claim on the cash flows from those mortgages. The pricing of these securities is influenced by factors such as interest rates, the underlying mortgage characteristics (for example, prepayment risk), and the overall demand for housing-related securities. Primary dealers are crucial in assessing and managing these risks, ensuring the securities are correctly priced for the market and providing liquidity to buyers and sellers alike.

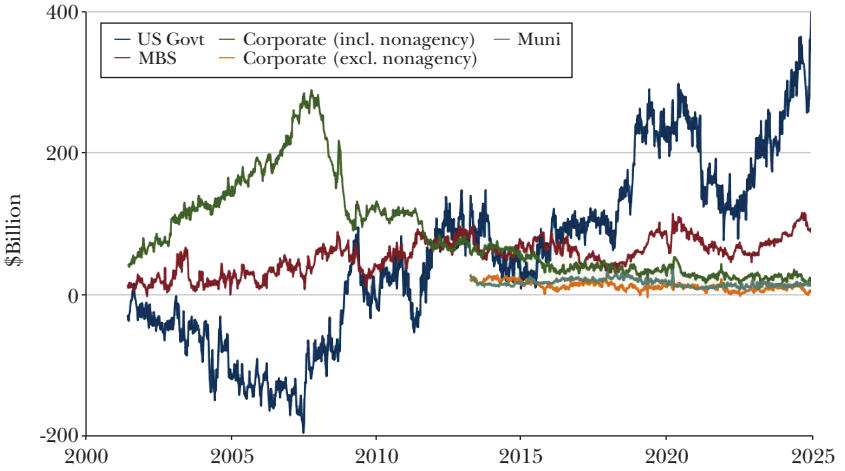
Municipal bonds are typically sold through either competitive bidding or negotiated underwriting. In competitive bidding, multiple underwriters bid to purchase the bond at the best price, and the issuer (local or state government) accepts the best bid. In negotiated underwriting, the issuer works directly with one or more underwriters to set the terms of the bond sale, including the price and the interest rate. The choice between these two methods depends on factors such as the size of the issue, the credit quality of the issuer, and market conditions. Larger, more complex deals are usually handled through negotiated underwriting, while smaller, less complex issues may be sold through competitive bidding.

Similarly, corporate bonds are usually sold in the primary market through negotiated underwriting, though in some cases, they may be sold through a private placement, where bonds are offered directly to a select group of institutional investors without a public offering. The negotiated underwriting process for corporate bonds often involves primary dealers, who assume the responsibility of pricing and distributing these securities to investors.

To some extent, one might argue that dealers play an even more significant role as intermediaries in the markets for corporate bonds, municipal bonds, and agency mortgage-backed securities, due to the heterogeneity of issuers and issues. These markets are characterized by a wide range of factors, such as varying amounts outstanding, maturities, call features, and other provisions. In addition, dealers assume significant credit risk in these markets, beyond the risk associated with changing interest rates, as they are responsible for intermediating securities with varying degrees of credit quality. This task adds complexity to their role, as they must assess and manage risks associated with the creditworthiness of issuers and the specific features of each security.

Figure 2 shows the net positions of primary dealers in US Treasuries, corporate bonds, municipal bonds, and agency mortgage-backed securities. While the sheer size of Treasuries' net positions now dominates the other fixed income markets, primarily driven by the large issuance of Treasuries since 2018, primary dealers both hold and intermediate substantial amounts in the other fixed income markets. As Figure 2 illustrates, in early 2000s, primary dealers were short Treasuries and long corporate and mortgage-backed securities. The negative net position of primary dealers in Treasuries means that they sold more than they bought. However, it is important to recognize that while primary dealers may not have held large long positions in Treasuries during that period, their role remained integral, especially in times of market stress when they acted as key liquidity providers. Their negative position likely reflected a combination of factors, such as rising interest rates, increased volatility, and regulatory changes

Figure 2
Primary Dealers' Net Positions in Fixed Income Securities, July 2001–June 2024



Source: Data are from the Federal Reserve Bank of New York's FR 2004A statistical release (https://www.newyorkfed.org/markets/counterparties/primary-dealers-statistics).
Note: This figure shows fixed income holdings of primary dealers as reported weekly in form FR-2004. It includes US Treasury securities, corporate bonds, mortgage-backed securities (MBS), and municipal bonds. Due to reporting changes and coverage additions, data on corporate debt positions are available from July 4, 2001, and municipal debt positions from April 3, 2013. The corporate debt series includes non-agency mortgage-backed securities from July 4, 2001, to April 3, 2013, so we show a corporate debt series that includes non-agency post-April 2013 (green), and another corporate bond series that excludes those securities once they are reported separately post-April 2013 (orange).

that made holding inventory more costly. Even when primary dealers held negative net positions, they continued their market-making and liquidity provision activities in Treasury auctions and secondary market transactions.

Although the data collection on positions of primary dealers does not allow us to directly compare the maturities of the corporate bonds and Treasuries that they hold, risk management consideration would suggest that Treasury and corporate bond positions should move in opposite directions, as Treasuries are used to hedge interest rate risk in the corporate bond market. In Figure 2, we see that that was indeed the case before 2007. However, the 2007–2009 financial crisis and the post-crisis regulatory changes affected dealers' incentives to hold inventory and provide intermediation in fixed income markets (for example, Adrian, Boyarchenko, and Shachar 2017), leading to an increase of Treasuries holdings to a positive territory and a decrease of corporate securities holdings.

Similar to the primary-secondary market pass-through observed in the US Treasury market, several studies have documented and quantified this phenomenon in other fixed income markets as well. This is not entirely surprising, as the obligation of primary dealers to participate in Treasury auctions can be seen as analogous to the commitments of dealers when underwriting corporate bonds

or municipal bonds, or when securitizing agency mortgage-backed securities. In looking at corporate bond markets, for example, Siani (2022) shows that the difference between primary and secondary market yields are countercyclical. Boyarchenko et al. (2024) show that the primary-secondary spread in the corporate bond market is positive and relatively small during "normal" periods, but it becomes negative—that is, the average price in the primary market is above the average price in the secondary market—and large during downturns. This pattern implies that primary market measures are necessary in addition to secondary market trading activity to capture credit conditions for nonfinancial borrowers. As we discuss below, signals from credit conditions measured on a broad basis—rather than just relying on secondary market prices—are a crucial input in determining when and how liquidity interventions should be conducted.

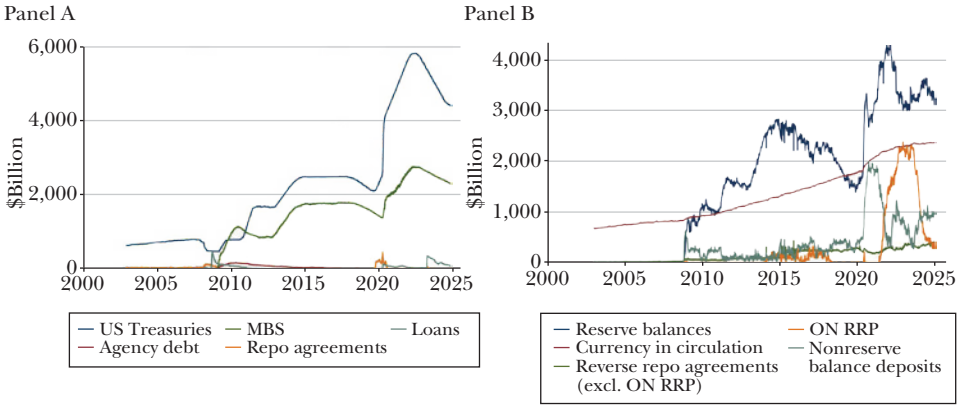
In the municipal bond market, Haughwout, Hyman, and Shachar (2021) show how Federal Reserve actions (discussed in more detail later) to support the municipal bond market during COVID-19 translates to improvements in the secondary market. In the market for mortgage-backed securities, Fuster et al. (2013) study the degree to which secondary market movements are reflected in mortgage borrowing costs by tracking cash flows during and after the mortgage origination and securitization process. More generally, the pass-through between primary and secondary markets depends on the health of the financial intermediaries common to both markets, which in the case of US fixed income markets are to a large extent primary dealers.

Fixed Income Markets and Monetary Policy Implementation

Monetary policy implementation in the United States has long relied on government-issued fixed income markets. In this section, we review the monetary policy implementation framework since the global financial crisis of 2007–2009, which focuses on policy implementation through changes in the size and composition of both the assets and the liabilities of the balance sheet of the central bank.

Historically, monetary policy implementation worked through controlling the overall quantity of reserves—money that member-banks deposit with the Federal Reserve—in the financial system by engaging in open market operations to adjust the amount of Treasury securities held by the central bank. In the pre-crisis monetary policy implementation regime, reserves were "scarce"—the total supply of reserves held at the Federal Reserve was relatively low—and changes in the size of Federal Reserve's assets translated directly into changes in the supply of reserves and thus the federal funds interest rate that clears (equilibrates the individual banks' demands for reserves with the total reserves supplied by the Federal Reserve) the market for reserves. The Federal Reserve thus has an implicit interest in ensuring that Treasury markets function smoothly so that the implementation and transmission of monetary policy remains effective. Primary (or otherwise designated) dealers have historically acted as counterparties to open market operations as these dealers act as large, significant intermediaries in both the primary and the secondary markets

Figure 3
The Federal Reserve Bank's Balance Sheet Composition



Source: Data are from Federal Reserve Statistical Release H.4.1(https://www.federalreserve.gov/RELEASES/H41/).
Note: This figure breaks down the composition of assets (panel A) and liabilities (panel B) of the Federal Reserve Bank's balance sheet. MBS stands for mortgage-backed securities; ON RRP stands for overnight reverse repurchase agreement.

for Treasury securities, ensuring that open market operations are executed at "fair prices" in desired volumes.

It is also worth noting that, even before the abundant reserves regime that has emerged since 2007–2009, the Federal Reserve recognized that reserves serve not only as a vehicle for the transmission of monetary policy but also act as a source of liquidity for the financial system. For example, in the aftermath of September 11, 2001, the Federal Open Market Committee both lowered the target rate and recognized the tension between maintaining the federal funds rate at the target level and supplying enough reserves to meet the demand for financial market settlements, with its September 17, 2001, statement: "The Federal Reserve will continue to supply unusually large volumes of liquidity to the financial markets, as needed, until more normal market functioning is restored. As a consequence, the FOMC recognizes that the actual federal funds rate may be below its target on occasion in these unusual circumstances." Indeed, the effective federal funds rate on September 18 and 19, 2001, was around 1.25 percent, while the federal funds target was 3 percent. The securities purchases necessary to create this unusually large supply of reserves highlight the key role that counterparties to open market operations play, especially during periods of stress in financial markets.

Figure 3 plots the evolution of the Federal Reserve's balance sheet over the last 25 years. Starting with panel A, which focuses on the assets held by the Federal Reserve, we see that the size of the central bank's balance sheet expanded substantially in the aftermath of the global financial crisis and again in response to the

COVID-19 pandemic in 2020. Furthermore, while the Federal Reserve primarily held US Treasuries before 2008, the composition of the balance sheet has since shifted to also include a substantial amount of agency mortgage-backed securities. On the liabilities side, shown in panel B, the post-crisis expansion in the Federal Reserve's balance sheet size was primarily funded by increases in reserve balances—that is, nondemandable deposits that member-banks hold with the Federal Reserve. In the aftermath of the COVID-19 pandemic, the overnight reverse repurchase facility, in which the Federal Reserve lends out securities on its balance sheet to market participants, also expanded.

So how does the Federal Reserve interact with fixed income market participants to implement monetary policy in the post-crisis monetary policy regime? While the federal funds rate was below its target for only a few days in September 2001, the global financial crisis of 2007–2009 and the associated market distress ushered in a new normal for monetary policy, one where reserves are either "abundant" or "ample" and therefore the size of the Federal Reserve's balance sheet plays a key role in the implementation of monetary policy. When reserves are "abundant" and the financial system is flush with reserves, changes in the effective federal funds rate have no impact on an individual bank's demand for reserves. When reserves instead are less than abundant but still "ample," there is still a large quantity of reserves available to the financial system but changes in the effective federal funds rate have a small impact on an individual bank's demand for reserves. Notably, as the amount of reserves supplied to the financial system declines, the economy moves from an abundant reserves regime to an ample reserves regime and, eventually, to the restrictive reserves regime under which conventional monetary policy historically operated.

Given that the willingness of banks to hold reserves in the abundant reserves regime is insensitive to the effective federal funds rate, how does the central bank control the level of interest rates? Since reaching the effectively zero percent lower bound for monetary policy in 2008, the stance of monetary policy has been communicated via a range for the federal funds rate, rather than a federal funds rate target. Operationally, the Federal Reserve maintains the effective federal funds rate within the monetary policy target by setting so-called administered rates; that is, rates that the Federal Reserve pays on its liabilities, with the Federal Reserve acting as the price setter in the corresponding markets.

The upper bound of the target range is maintained via the interest rate paid by the Federal Reserve on bank reserve balances, and captures the trade-off that a bank faces between depositing reserves at the Federal Reserve overnight, earning the interest rate on reserves balances, or lending in other markets or purchasing securities to hold overnight. The counterparties to maintaining the effective federal funds rate below the upper bound of the target range are thus the same as the counterparties to maintaining the federal funds rate at the policy target in the restrictive reserves regime: the Federal Reserve member-banks.

Just like the upper bound of the target range captures the trade-off that a bank faces, the lower bound of the target range captures the trade-off between lending to the Federal Reserve at an administered rate and lending in the private market for

a broader set of market participants. In particular, the Federal Reserve maintains a standing (that is, operating regardless of market conditions) overnight reverse repo facility. Counterparties in the reverse repo facility face the trade-off of lending at the overnight reverse repo rate to the Federal Reserve or, once again, lending in other markets or purchasing securities to hold overnight. The overnight reverse repo facility has an expanded set of counterparties relative to those that have access to deposit accounts at the Federal Reserve and includes, notably, money market funds, as well as government-sponsored enterprises, primary dealers, and other expanded counterparties. By expanding the access to the overnight reverse repo facility to include money market funds, the stance of monetary policy is transmitted directly to an important repository of liquidity in the modern financial system. However, in expanding the list of counterparties, the Federal Reserve follows the same principles as in designating primary dealers. These new counterparties are chosen after an evaluation of their capabilities, including having pre-existing repo arrangements in the relevant markets, and have minimum participation requirements.[5]

To summarize, in the monetary policy implementation regime since the global financial crisis of 2007–2009, the liabilities of the central bank include a large (abundant or ample) level of reserves and borrowings from the overnight reverse repo facility. By controlling the size of the Federal Reserve balance sheet through asset purchases, the Federal Reserve thus controls the overall amount of liquidity supplied to the financial system through reserves and the size of the overnight reverse repo facility.

Under this new normal for monetary policy, not just the size but also the composition of the Federal Reserve's balance sheet and the pace and timing of purchases (expansions of the balance sheet) and of balance sheet run-off all play a role. In particular, the Federal Reserve's balance sheet can also be used to affect market conditions directly for policy easing purposes, as well as to stabilize markets during periods of stress by providing an alternative long-term holder of securities. The Federal Reserve's large-scale asset purchases began in 2008, at the depth of the global financial crisis, with the Federal Open Market Committee announcing on November 25, 2008, that the Federal Reserve would begin purchases of agency debt and agency mortgage-backed securities "to reduce the cost and increase the availability of credit for the purchase of houses, which in turn should support housing markets and foster improved conditions in financial markets more generally." While this initial round of "quantitative easing"—expanding the monetary base in the economy through an expansion of the Federal Reserve's balance sheet—was conducted for market stabilization purposes, subsequent rounds of quantitative easing were targeted towards easing the overall stance of monetary policy and thereby a relaxation of financial conditions by reducing long-term rates in the economy beyond the levels implied by the federal funds target range.

[5] For a full list of overnight reverse repo counterparty requirements, see https://www.newyorkfed.org/markets/rrp_counterparties.

Quantitative easing purchases are mainly conducted using primary dealer counterparties, just like historical open market operations. However, the view of the role of transactions with primary dealers is somewhat different under balance sheet policy. Although primary dealers still serve the market-making role, enabling the Federal Reserve to conduct purchases at scale at fair prices, one of the channels of transmission of monetary policy is through the relaxation of so-called balance sheet constraints at primary dealers. Balance sheet constraints convey the idea that financial institutions, either because of risk-management motives, regulation, or market forces, face an overall constraint on the size of their balance sheets, as well as potentially on the riskiness of their balance sheets. During periods of market stress, balance sheet constraints are more likely to bind, as market volatility may affect the permissible level of risk taken on by financial institutions, thereby constraining asset holdings.

Central bank purchases thus alleviate balance sheet constraints in three ways. First, central bank purchases, by expanding the size of the central bank balance sheet, provide additional liquidity to the financial system, easing constraints. In other words, expanding the size of the balance sheet expands the supply of reserves to the financial system, providing space on member-banks' balance sheets for greater borrowing through deposits. This is the same channel that the Federal Open Market Committee contemplated in providing an elevated level of reserves during September 2001, as we discussed above.

Second, central bank purchases remove securities from the balance sheets of primary dealers. This allows primary dealers to invest in other securities, restarting the two-sided market making in which dealers are normally engaged. The quantitative easing programs in 2008 and in the wake of the COVID-19 pandemic in 2020 were designed to operate in this fashion, removing securities from dealer balance sheets and stimulating primary dealers' intermediation activity not just in the directly affected markets but in all fixed income markets. Likewise, the maturity extension program in 2011–2012, which kept the size of the Federal Reserve's balance sheet constant but shifted the composition toward longer-term securities by financing purchases of longer-term Treasury securities with sales of shorter-term Treasuries, aimed to reduce long-term rates in the economy by removing longer-term Treasuries from private balance sheets.

Third, Federal Reserve purchases reduce market volatility, further relaxing balance sheet constraints at financial institutions. By communicating a pace of purchases, central bank purchases provide predictable demand for securities. That is, during quantitative easing, the central bank commits to buying (at least) a certain quantity of securities in the secondary markets every month, reducing the need for distressed institutions to liquidate their positions at lower prices. Moreover, by decreasing long-term interest rates, purchases also reinforce the forward guidance provided by the Federal Open Market Committee, further stabilizing interest rate volatility.

It is worth emphasizing that the nature of quantitative easing has evolved substantially since the first announcement in November 2008. While the initial quantitative easing program was limited in scope, and even initially excluded US Treasuries, the

size of the subsequent programs, the types of securities purchased, and the pace of purchases were materially expanded. Of particular interest are the quantitative easing purchases in 2020, which were unprecedented in their speed and scale and innovative in their implementation. On March 23, 2020, the Federal Open Market Committee announced that it would continue to purchase securities "in the amounts needed" to support market functioning and effective transmission of monetary policy. This allowed the Federal Reserve Bank of New York to scale up purchases quickly and to vary the pace and distribution of purchases based on observable measures of market functioning. During the height of the dislocations in spring 2020, purchases reached $100 billion per day, and totaled more than $2 trillion by April 30, 2020 (Fleming et al. 2022). Furthermore, the Federal Reserve purchased for the first-time agency *commercial* mortgage-backed securities, adjusting once again—just as it did in 2008—to the particular sources of stress to the financial system.

Our discussion so far has focused on how monetary policy easing operates in the abundant reserves regime. Unlike monetary policy in the restricted reserves regime, the implementation of monetary policy tightening in the abundant reserves regime is not a simple reversal of the implementation of easing policies. Because monetary policy tightening begins in the abundant reserves regime, where the quantity of reserves is essentially insensitive to small changes in the effective federal funds rate, a key component of monetary policy tightening is a reduction of the size of the Federal Reserve's balance sheet. The Federal Reserve has so far implemented reductions in balance sheet size through reductions in reinvestments rather than outright sales.[6] By communicating the path of reductions in reinvestments, the Federal Open Market Committee keeps quantitative tightening predictable, and thus does not undo the reductions in uncertainty achieved through quantitative easing. Reducing balance sheet size by capping reinvestments instead of outright sales also means that financial institutions only have to absorb progressively more of new issuances of Treasury securities and agency debt and mortgage-backed securities, rather than absorbing the stock of securities held by the Federal Reserve. This reduces the potential stress to the balance sheets of primary dealers during quantitative tightening. That is, while primary dealers are counterparties to increasing the size of the Federal Reserve's balance sheet, they are not direct counterparties to reducing the amount of assets held by the Federal Reserve.

Instead, reductions in the size of the Federal Reserve's balance sheet have a tightening effect primarily by removing liquidity from the financial system (Smith and Valcarcel 2023). Because the liabilities of the Federal Reserve in the current implementation of monetary policy include both reserves and borrowings from the overnight reverse repo facility, the counterparties to quantitative tightening are thus

[6] While easing monetary policy (expansions in the size of the balance sheet) or holding steady the stance of monetary policy (keeping the size of the balance sheet constant), the Federal Reserve reinvests the principal repayments—that is, the amount it receives when a Treasury security reaches maturity or when a mortgage-backed security is prepaid or reaches maturity into new securities. This prevents declines in the size of the balance sheet for mechanical reasons rather than active policy choices.

member banks and money market funds. Because the overnight reverse repo rate is the lower bound of the effective federal funds rate, monetary policy tightening first leads to reductions in the average level of borrowings from the overnight reverse repo facility, as money market funds seek alternative investments to meet their interest rate obligations. Reserves, however, are more sticky as banks have historically passed through increases in the target range to deposit rates to a much lower extent than money market funds have passed through increases in the target range to money market interest rates. Indeed, during the current balance sheet reduction program which began in June 2022, reduction in reserve balances did not begin until the borrowings in the overnight reverse repo facility almost reached zero.

Fixed Income Markets and Liquidity Interventions

In the previous section, we discussed how the Federal Reserve system interacts with market participants and, in particular, primary dealers for the purposes of implementing monetary policy. As such, the Federal Reserve has an inherent interest in the smooth functioning of markets that it uses to implement monetary policy, namely that for US Treasury and agency debt and mortgage-backed securities. Beyond monetary policy implementation, the Federal Reserve has an interest in smoothly functioning financial markets for a number of reasons. In a well-functioning market, the market-clearing asset prices accurately reflect market participant beliefs. For example, a pessimistic outlook on real activity would translate into lower Treasury market yields, a decreased confidence in the health of productive firms into higher corporate bonds yields, and expectations of higher future inflation into higher yields of inflation-linked Treasuries. The Federal Reserve may then use asset prices and financial conditions more broadly to learn about market participants' beliefs, including those about the future path of inflation, the future path of unemployment, or the future path of monetary policy.

Financial markets also serve to channel credit to different parts of the economy. Well-functioning fixed income markets are particularly important in this regard, with firms using the commercial paper and corporate bond markets as alternative sources of credit to bank loans, households relying on an active market in mortgage-backed securities for mortgage origination, and the federal and state governments relying on Treasury and municipal markets, respectively, to finance their debt. As such, Section 13(3) of the Federal Reserve Act (added through the Emergency Relief and Construction Act of 1932) allows the Federal Reserve "in unusual and exigent" circumstances to lend to individuals, partnerships, and corporations, provided that the participants in such programs are "unable to secure adequate credit accommodations from other banking institutions." Before discussing the particular 13(3) facilities—so-called because they were created using the authority provided by Section 13(3) of the Federal Reserve Act—that were implemented during the global financial crisis and in response to market

disruptions during the COVID-19 pandemic, it is important to emphasize several features of such facilities.

First, Section 13(3) provides the Federal Reserve with authority to lend to nonbank borrowers. Section 13(3) thus expands the "lender of last resort" mandate of the central bank beyond the banking sector when the circumstances warrant such interventions. Member banks of the Federal Reserve instead have access to the discount window, which allows banks to borrow from the Federal Reserve against collateral.

Second, lending conducted under the authority of Section 13(3) is limited to "unusual and exigent" circumstances, so that 13(3) facilities can only be used for market stabilization purposes, not for monetary policy. In other words, a 13(3) facility cannot be created to, for example, lower long-term interest rates further when the monetary policy target rate is at the effective zero lower bound. As such, 13(3) facilities require an authorizing vote of at least five members of the Federal Reserve Board of Governors. The Wall Street Reform and Consumer Protection Act of 2010, commonly called the Dodd-Frank Act, further requires the Federal Reserve to obtain prior approval from the Secretary of the Treasury before establishing any lending facility under Section 13(3), and prohibits loans to individual firms except through programs that are broadly available to many firms. These provisions act to ensure that 13(3) facilities are only used during periods of broad market stress.

Third, because 13(3) facilities are used to address market dysfunction, 13(3) facilities are in operation for a limited time. Even during the Great Depression, lending under Section 13(3) was reauthorized for six-month intervals from July 26, 1932 to July 31, 1936. Similarly, the last of the facilities put in place during the global financial crisis of 2007–2009 stopped lending at the end of December 2010, and the last of the COVID-19 facilities stopped extending new loans at the end of December 2020. Thus, unlike large-scale asset purchases for the purpose of monetary policy implementation, which proceed at a prespecified pace until monetary policy objectives have been reached, the making of new loans through 13(3) facilities are authorized for short periods of time only (though the operational period of the facilities may be extended if market conditions remain strained).

Finally, only creditworthy borrowers can borrow from 13(3) facilities. By lending to only creditworthy borrowers, 13(3) facilities limit the credit risk exposure of the facilities, limit redistribution of credit—lending to borrowers who would not be able to borrow under normal market conditions—and limit the moral hazard concerns of extending (emergency) lending to a broader set of borrowers. The usual funding structure for 13(3) facilities includes an amount of equity financing provided by the Federal Reserve Bank standing up the facility, which is leveraged to the overall size of the facility using loans from the US Treasury. Losses on loans made by 13(3) facilities are thus first borne directly by the US Treasury and, only if such losses were to exceed the funding initially provided by the US Treasury, by the Federal Reserve Bank. Historically, no 13(3) facility has generated losses. Any net revenue generated by 13(3) facility loans is rebated to the US Treasury.

How do 13(3) facilities limit the credit risk of loans extended, even while expanding the set of borrowers who have access to emergency Federal Reserve

lending and allowing for non-recourse loans?[7] Facilities that extend loans directly usually focus on securing adequate collateral. For example, the Federal Reserve may require specific types of collateral, with high ratings from the Nationally Recognized Statistical Rating Organizations (NRSROs), with the value of collateral "haircut" relative to its face value; that is, such facilities lend less than the full value of collateral pledged to secure the loan, and require that the collateral is of high quality. Furthermore, facilities charge above-normal market interest rates and usage fees to ensure that facility borrowing only occurs when markets are not functioning properly, and engage with specialized vendors to obtain expertise to perform critical functions within the lending facilities. Similarly, facilities that buy securities in the secondary market require ratings from the NRSROs, bid for securities at prices below normal market prices to ensure that the facilities purchase securities only when markets are not functioning properly, and engage with both specialized fiscal agents and authorized counterparties in implementing the secondary market transactions.

Before describing the full set of 13(3) facilities used to stabilize access to credit during the global financial crisis and the COVID-19 pandemic, it is worth illustrating how 13(3) facilities operate using a particular example. The onset of the COVID-19 pandemic brought about widespread distress in global financial markets, including the US corporate bond market. Spreads on corporate bonds rose to levels not seen since the global financial crisis, issuance of new corporate bonds slowed, and liquidity in secondary markets deteriorated to levels even lower than during the global financial crisis.[8]

In response to the "unusual and exigent" circumstances of the COVID-19 pandemic, the Federal Reserve announced on March 22, 2020, the creation of Primary and Secondary Market Corporate Credit Facilities (PMCCF, SMCCF; jointly referred to as the CCFs). Borrower credit worthiness for these facilities was evaluated based on borrower credit rating, with only "investment grade" (those whose credit rating is sufficiently high) and "fallen angel" (those borrowers who were investment grade prior to the onset of the pandemic) borrowers permitted to borrow from the facilities. As discussed in greater detail in Boyarchenko et al. (2020) and Boyarchenko, Kovner, and Shachar (2020), the CCFs supported access to credit for borrowers who, at the time, employed more than 16 million people and whose bonds are key assets for retirees and pension funds. The announcement of the facilities on March 22, 2020, brought a rapid normalization of bond credit spreads and restarted issuance of new corporate bonds, with total issuance in April 2020 one of the highest on record.

Liquidity Facilities to Support Particular Types of Institutions

Broadly speaking, 13(3) facilities that were implemented during the global financial crisis and during the market disruptions associated with the

[7] Non-recourse loans are those that do not allow the lender access to any assets of the borrower not pledged as collateral in the loan.

[8] See, for example, Boyarchenko et al. (2020) for a real-time description of the case for interventions in the US corporate bond markets in 2020.

COVID-19 pandemic can be split into two categories. One category, which we will tackle first, are those facilities that support the activities of particular types of institutions, deemed critical to the stable functioning of markets in general. The other category, which we will discuss in more detail in the next subsection, consists of facilities that support activity in particular markets.

Both during the global financial crisis and the market disruptions associated with the COVID-19 pandemic, 13(3) facilities targeted financial institutions central to broad financial markets. The first of such facilities is the Primary Dealer Credit Facility (PDCF), which provided funding to primary dealers with the purpose of supporting market liquidity and functioning and facilitating credit availability to businesses and households. The PDCF—first stood up on March 16, 2008, in response to the bail-out of Bear Stearns, and reauthorized on March 17, 2020, in response to the COVID-19 pandemic—acted as the equivalent of discount window lending for primary dealers. In particular, PDCF provided collateralized loans to primary dealers, with the interest rate set equal to the discount window's primary credit rate and the haircut and margins applied to the posted collateral determined in a similar manner to those applied to collateral posted in discount window loans. Aside from giving access to discount-window-like loans to institutions (primary dealers) that do not have access to the discount window, the PDCF also expanded the set of eligible collateral (for example, including equities) to include security types not normally held by banks.[9] Furthermore, the 2020 PDCF allowed for term loans (up to a maximum of 90 days), while the 2008 PDCF allowed for overnight loans only.

Because in both 2008 and 2020 the Primary Dealer Credit Facility was intended to help the repo market to continue functioning efficiently during an adverse liquidity spiral, the funding provided through the PDCF took the form of repurchase agreements that settled through triparty repo. In a triparty repo agreement, the borrower (primary dealer in the case of PDCF) posts collateral with the clearing bank and receives cash from the lender (the PDCF). In 2020, the authorization of the PDCF served to normalize repo spreads on PDCF-eligible collateral quickly. In 2008, though the PDCF was successful as a liquidity back-stop to primary dealers and repo markets in the immediate aftermath of the liquidation of Bear Stearns, the failure of Lehman Brothers in September 2008 brought further stress to the market, and the 2008 PDCF was expanded from only accepting collateral eligible for discount window loans to the broader set of collateral used in triparty repo agreements. Overall, the usage of the 2020 facility peaked at around $40 billion in April 2020, while the usage of the 2008 facility peaked at around $140 billion in September 2008.

The 2008 Primary Dealer Credit Facility was complemented by the Term Securities Lending Facility (TSLF), which loaned Treasury securities to primary dealers

[9] See Martin and McLaughlin (2022) for a detailed discussion of the 2020 Primary Dealer Credit Facility and a comparison to the 2008 PDCF, and Adrian, Burke, and McAndrews (2009) for a detailed discussion of the 2008 PDCF.

against eligible collateral, thereby providing dealers with higher-quality collateral to use in private market funding transactions, promoting liquidity in US Treasury markets and markets for other assets.

The other financial sector deemed key for the continuing functioning of asset markets during both the global financial crisis and the COVID-19 pandemic was the money market mutual fund sector. During both episodes of market stress, 13(3) facilities to assist money market funds in meeting demands for redemption by their clients (households in particular) were authorized. The 2008 Asset-Backed Commercial Paper Money Market Liquidity Facility (AMLF) provided financing to financial institutions for purchases only of asset-backed commercial paper from money market mutual funds. The 2020 Money Market Mutual Funds Liquidity Facility (MMLF) allowed for purchases of a broader set of assets from money market funds, more closely aligning with the assets held by money market funds.[10] In both cases, the money market fund facilities were meant to stop runs on money market funds by providing a backstop to the value of (high quality) securities held by money market funds. The 2008 AMLF facility was further supplemented with the Money Market Fund Investor Funding Facility (MMIFF). Just as AMLF provided funding to purchase asset-backed commercial paper from money market mutual funds, the MMIFF provided funding to purchase certificates of deposit and commercial paper issued by highly rated financial institutions.

To summarize, both the primary dealer and the money market mutual fund 13(3) facilities were implemented to support—through maintaining the flow of funding to primary dealers in the repo market and through limiting fire-sale redemptions at money market funds—the crucial market making activity conducted by these financial institutions. In other words, these facilities served to maintain access to funding for financial intermediaries and, as such, fundamentally served a role that cannot be accomplished through monetary policy lowering economy-wide interest rates alone.

Liquidity Facilities to Support Particular Markets

In addition to supporting access to credit for key financial institutions, 13(3) facilities were established to support specific markets directly. During the global financial crisis, disruptions in financial markets stemmed from distress at financial institutions. As such, the 13(3) facilities established during the global financial crisis focused on markets in which intermediation was historically conducted by affected institutions: the commercial paper market and the asset-backed securities market. Both the Commercial Paper Funding Facility (CPFF) and the Term Asset-Backed Securities Loan Facility (TALF) facilitated primary market issuance in their respective markets. The CPFF thus supported access to credit for highly-rated firms, while TALF supported access to credit for households and businesses through instruments (including credit cards and automobile loans) that rely on a functioning

[10] See Anadu et al. (2022) for an analysis of the Money Market Mutual Funds Liquidity Facility.

asset-backed securities market. The peak utilization of the 2008 CPFF facility was $348.2 billion in January 2009 and the peak utilization of the 2008 TALF was $48.2 billion in March 2010.

The market disruptions associated with the COVID-19 pandemic were much broader and stemmed as much from uncertainty about the ability of nonfinancial borrowers to remain solvent over the (uncertain) duration of the pandemic as from stress at financial institutions (as was the case during the global financial crisis). Thus, although the Commercial Paper Funding Facility and Term Asset-Backed Securities Loan Facility were also reauthorized during March 2020, a much broader set of liquidity facilities was necessary.[11] COVID-19 period 13(3) facilities were expanded to include support to nonfinancial corporate bond borrowers (primary and secondary market corporate credit facilities, PM/SMCCF), support to small and medium-sized businesses (Main Street Lending Facilities, MSLF), municipalities (Municipal Liquidity Facility), and paycheck lending to smallest businesses (Paycheck Protection Program Liquidity Facility, PPPLF).[12]

The corporate credit facilities, Main Street lending facilities, and the municipal liquidity facility deviated from the design of the Commercial Paper Funding Facility and the Term Asset-Backed Securities Loan Facility by including a secondary market component to the facilities. The Secondary Market Corporate Credit Facilities and the municipal liquidity facility purchased eligible securities in the secondary market, while the Main Street Expanded Loan Facility was authorized to purchase up to 95 percent of an upsized tranche of eligible, pre-existing term or revolving credit loans. What was the motivation behind including a secondary market element in these facilities? When there exists an active secondary market, such as is the case for corporate and municipal bonds and for syndicated loans, market conditions in the secondary market spill over into market conditions for issuers. Such spillovers can occur because primary market prices are set as a function of secondary market prices on comparable instruments. Moreover, a liquid secondary market provides insurance to primary market lenders in case the primary market lender would like to resell existing positions. Boyarchenko, Kovner, and Shachar (2022) explore the primary-secondary market interconnections in the corporate bond market, and show the differential effects of the corporate credit facilities on primary and secondary market intermediaries.

Where to from Here?

Fixed income markets serve as a crucial link between the financial system and the real economy, facilitating capital allocation, government funding, and

[11] For a description of the 2020 Commercial Paper Funding Facility and the Term Asset-Backed Securities Loan Facility CPFF, see Boyarchenko et al. (2022b) and Caviness et al. (2022).

[12] For detailed descriptions and analysis of these facilities, see Boyarchenko et al. (2022a), Arseneau et al. (2022), Haughwout, Hyman, and Shachar (2022), and Volker (2022), respectively.

corporate financing. The efficiency and stability of these markets directly impact borrowing costs, investment decisions, and overall economic growth. Because the Federal Reserve has historically relied on banks and primary dealers as the first link in the chain to transmit its policies to the broader economy, our essay focused on their key role in supporting market liquidity and maintaining well-functioning fixed income markets. As we reviewed earlier, the landscape for fixed income ownership has shifted after the 2007–2009 financial crisis, and then again after March 2020 crisis. The composition of investors in fixed income markets is now more dispersed (Figure 1), and it is no longer clear who is the marginal investor in each of those markets. Looking forward, the interlinks between primary dealers and non-bank financial institutions as issuers, end users, and intermediaries should be better understood, especially in the context of conventional and unconventional monetary policy pass-through.

Non-bank financial institutions—such as mutual funds, money market funds, exchange-traded funds, pension funds, insurance companies, and government-sponsored enterprises—are financial companies that do not have a banking license. As of the end of 2024, they are more than three times larger than the US banking system.[13] They have been, in part, substituting dealers' diminished willingness to intermediate and allocate capital for trading fixed income securities, and improve market liquidity (for example, Dannhauser 2017). Nonetheless, non-bank financial institutions are subject to lighter reporting requirements and regulations, allowing them to take on more leverage than banks but without required capital buffers and access to emergency liquidity programs. Moreover, while they might play a crucial role in liquidity provision in fixed income markets during normal times, their exit in stress times might amplify shocks, even if they are not the source of the shock. Indeed, in March 2020, non-bank financial institutions' presence in Treasuries, agency mortgage-backed securities, corporate bonds, and municipal bonds markets played a critical role propagating the "dash for cash" across those markets (for example, Haddad, Moreira, and Muir 2021; Ma, Xiao, and Zeng 2022).

The growing importance of non-bank financial institutions is not only a financial stability issue, but also raises new questions about monetary policy implementation. The participation of investment funds—including mutual funds, money market funds, hedge funds, money managers, and investment advisors—in auctions of Treasury securities increased from 1.7 percent in January 2008 to 67.8 percent in October 2023, whereas the share attributable to dealers and brokers' share decreased from 79 percent to 19.4 percent during the same period. The counterparties of the Fed's open market operations in the secondary market, however, still rely exclusively on primary dealers.

Expanding the eligibility criteria for the Fed's Treasury operations to include non-bank financial institutions could improve market liquidity, particularly during market disruptions. Similar to non-bank financial institutions' participation in the

[13]This is based on data from Financial Accounts of the United States (available at https://www.federalreserve.gov/releases/z1/).

primary market, allowing a broader mix of participants to sell Treasuries to the Fed directly could provide liquidity to the market more directly and immediately, thereby supporting the pass-through of the Fed's policy. Moreover, a broader mix of participants in the Fed's operations could lead to more competitive pricing for the Fed.

Future research and policy should account for the heterogeneity of non-bank financial institutions, including differences in risk appetite, duration preferences, accounting standards, and regulatory and internal constraints. These variations can contribute to segmentation both across and within fixed income markets, influencing liquidity dynamics and market stability. Although the Federal Reserve has effectively been using its available policy toolkit to address liquidity disruptions during recent recessions, fixed income markets have also experienced abrupt episodes of liquidity deteriorations—such as the "Taper Tantrum" on May 2013 and the "Flash Rally" on October 15, 2014—where non-bank financial institutions played a role in amplifying volatility. To better understand and mitigate such risks, comprehensive data collections and enhanced transparency will be essential.

■ *The views expressed here are the authors' and are not representative of the views of the Federal Reserve Bank of New York or the Federal Reserve System. The authors are grateful for research assistance from Allen Liu, and for detailed comments from Michael Fleming and from the editors of the Journal of Economic Perspectives, Jonathan Parker, Nina Pavcnik, and Timothy Taylor.*

References

Adrian, Tobias, Nina Boyarchenko, and Or Shachar. 2017. "Dealer Balance Sheets and Bond Liquidity Provision." *Journal of Monetary Economics* 89: 92–109.

Adrian, Tobias, Christopher R. Burke, and James McAndrews. 2009. "The Federal Reserve's Primary Dealer Credit Facility." *Current Issues in Economics and Finance* 15 (4): 1–11.

Anadu, Kenechukwu, Marco Cipriani, Ryan Craver, and Gabriele La Spada. 2022. "The Money Market Mutual Fund Liquidity Facility." *Economic Policy Review* 28 (1): 139–60.

Arseneau, David, José Fillat, Molly Mahar, Donald Morgan, and Skander Van den Heuvel. 2022. "The Main Street Lending Program." *Economic Policy Review* 28 (1): 58–92.

Boyarchenko, Nina, Caren Cox, Richard K. Crump, Andrew Danzig, Anna Kovner, Or Shachar, and Patrick Steiner. 2022a. "The Primary and Secondary Corporate Credit Facilities." *Economic Policy Review* 28 (1): 1–34.

Boyarchenko, Nina, Richard K. Crump, Anna Kovner, and Deborah Leonard. 2022b. "The Commercial Paper Funding Facility." *Economic Policy Review* 28 (1): 114–29.

Boyarchenko, Nina, Richard K. Crump, Anna Kovner, and Or Shachar. 2024. "Measuring corporate Bond Market Dislocations." Federal Reserve Bank of New York Staff Report 957.

Boyarchenko, Nina, Richard K. Crump, Anna Kovner, Or Shachar, and Peter Van Tassel. 2020. "The

Primary and Secondary Market Corporate Credit Facilities." *Liberty Street Economics*, May 26. https://libertystreeteconomics.newyorkfed.org/202%5/the-primary-and-secondary-market-corporate-credit-facilities/.

Boyarchenko, Nina, Anna Kovner, and Or Shachar. 2020. "The Impact of the Corporate Credit Facilities." *Liberty Street Economics*, October 1. https://libertystreeteconomics.newyorkfed.org/2020/10/the-impact-of-the-corporate-credit-facilities/.

Boyarchenko, Nina, Anna Kovner, and Or Shachar. 2022. "It's What You Say and What You Buy: A Holistic Evaluation of the Corporate Credit Facilities." *Journal of Financial Economics* 144 (3): 695–731.

Brain, Doug, Michiel De Pooter, Dobrislav Dobrev, Michael Fleming, Peter Johansson, Frank Keane, Michael Puglia, Tony Rodrigues, and Or Shachar. 2018. "Breaking Down TRACE Volumes Further." FEDS Notes, November 29. https://www.federalreserve.gov/econres/notes/feds-notes/breaking-down-trace-volumes-further-20181129.html.

Caviness, Elizabeth, Asani Sarkar, Ankur Goyal, and Woojung Park. 2022. "The Term Asset-Backed Securities Loan Facility." *Economic Policy Review* 28 (1): 161–84.

Dannhauser, Caitlin D. 2017. "The Impact of Innovation: Evidence from Corporate Bond Exchange-Traded Funds (ETFs)." *Journal of Financial Economics* 125 (3): 537–60.

Fleming, Michael, Giang Nguyen, and Joshua Rosenberg. 2024. "How Do Treasury Dealers Manage Their Positions?" *Journal of Financial Economics* 158: 103885.

Fleming, Michael J., Haoyang Liu, Rich Podjasek, and Jake Schurmeier. 2022. "The Federal Reserve's Market Functioning Purchases." *Economic Policy Review* 28 (1): 210–41.

Fuster, Andreas, Laurie S. Goodman, David O. Lucca, Laurel Madar, Linsey Molloy, and Paul Willen. 2013. "The Rising Gap between Primary and Secondary Mortgage Rates." *Economic Policy Review* 19 (2): 17–39.

Garbade, Kenneth D. 2004. "The Institutionalization of Treasury Note and Bond Auctions, 1970–75." *Economic Policy Review*: 29–45.

Garbade, Kenneth D. 2016. "The Early Years of the Primary Dealer System." Federal Reserve Bank of New York Staff Report 777.

Haddad, Valentin, Alan Moreira, and Tyler Muir. 2021. "When Selling Becomes Viral: Disruptions in Debt Markets in the COVID-19 Crisis and the Fed's Response." *Review of Financial Studies* 34 (11): 5309–51.

Haughwout, Andrew, Benjamin Hyman, and Or Shachar. 2021. "The Option Value of Municipal Liquidity: Evidence from Federal Lending Cutoffs during COVID-19." Federal Reserve Bank of New York Staff Report 988.

Haughwout, Andrew, Benjamin Hyman, and Or Shachar. 2022. "The Municipal Liquidity Facility." *Economic Policy Review* 28 (1): 35–57.

Hortaçsu, Ali, Jakub Kastl, and Allen Zhang. 2018. "Bid Shading and Bidder Surplus in the US Treasury Auction System." *American Economic Review* 108 (1): 147–69.

Jegadeesh, Narasimhan. 1993. "Treasury Auction Bids and the Salomon Squeeze." *Journal of Finance* 48 (4): 1403–19.

Lou, Dong, Hongjun Yan, and Jinfan Zhang. 2013. "Anticipated and Repeated Shocks in Liquid Markets." *Review of Financial Studies* 26 (8): 1891–912.

Ma, Yiming, Kairong Xiao, and Yao Zeng. 2022. "Mutual Fund Liquidity Transformation and Reverse Flight to Liquidity." *Review of Financial Studies* 35 (10): 4674–711.

Malvey, Paul F., and Christine M. Archibald. 1998. *Uniform-Price Auctions: Update of the Treasury Experience.* Office of Market Finance, US Department of the Treasury.

Martin, Antoine, and Susan McLaughlin. 2022. "The Primary Dealer Credit Facility." *Economic Policy Review* 28 (1): 130–38.

Potter, Simon. 2015. "Challenges Posed by the Evolution of the Treasury Market." Speech, Primary Dealer Meeting at the Federal Reserve Bank of New York, NY, April 13, 2015. https://www.newyorkfed.org/newsevents/speeches/2015/pot150413.

Siani, Kerry. 2022. "Raising Bond Capital in Segmented Markets." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4239841.

Smith, A. Lee, and Victor J. Valcarcel. 2023. "The Financial Market Effects of Unwinding the Federal Reserve's Balance Sheet." *Journal of Economic Dynamics and Control* 146: 104582.

Volker, Desi. 2022. "The Paycheck Protection Program Liquidity Facility." *Economic Policy Review* 28 (1): 185–209.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 195–214

How US Treasuries Can Remain the World's Safe Haven

Darrell Duffie

US Treasury securities are, by far, the first choice of safe-haven investors around the world. Treasuries serve two distinct safe-haven roles. First, Treasuries are a crisis hedge. For example, over half of the world's official foreign-exchange reserves are invested in US Treasuries. As the risk of a crisis rises, investors seeking lower risk buy Treasuries in a "flight to quality." For instance, as the risk of a global pandemic rose in the weeks leading up to March 2020, a flight to quality increased the prices of US Treasuries, even relative to other safe government securities (Barone et al. 2022).

The second safe-haven role of Treasuries is a store of value that can be quickly liquidated for cash once a crisis actually hits. For this, the market must be able to quickly intermediate huge volumes of demand by investors to *sell* Treasuries. When the World Health Organization declared COVID-19 a global pandemic on March 12, 2020, the Treasury market was unable to meet this second safe-haven role effectively because the dealers who make markets for Treasuries were unable to handle the flood of demands by investors around the world to buy their Treasury securities. Bond dealers were asked at the same time to buy enormous quantities of mortgage-backed securities and corporate bonds, among other demands for liquidity. Total customer-to-dealer bond-market trade volumes suddenly jumped

■ *Darrell Duffie is Adams Distinguished Professor of Management and Professor of Finance, Graduate School of Business, and Senior Fellow, Stanford Institute for Economic Policy Research, both at Stanford University, Stanford, California. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is duffie@stanford.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241412. The author is a trustee of an organization whose interests relate to the article and has received financial support from that organization.

to over ten times their respective 2017–2022 sample medians (Duffie et al. 2023). The bond market reached the limits of its intermediation capacity and became effectively dysfunctional. Yields for Treasury securities lurched higher, while dealer-to-customer bid-offer spreads and dealer-to-dealer market depth worsened by factors of over ten (Duffie 2020). Among other steps to support the market, the Federal Reserve purchased almost $1 trillion dollars of Treasury securities from primary dealers in the first three weeks after March 12, freeing dealer balance-sheet space to handle more sales from customers. Weak market functionality persisted for several additional weeks (Duffie et al. 2023). Although liquidity in Treasury markets gradually returned to normal, many Treasuries investors presumably noticed that in the heart of the March 2020 crisis, they had not benefited from the safe-haven requirement of a liquid and deep market.

Even before the COVID-19 crisis, the vaunted liquidity of the market for trading Treasuries had been showing cracks under stress. As the quantity of US federal debt has reached the level of GDP for the first time since World War II, and continues to rise rapidly, there is a growing imbalance between the supply of Treasury securities and the peak intermediation capacity of Treasuries dealers. It would be premature to claim that the safe-haven status of Treasuries is imperiled by an exhaustion of US fiscal space. Bonds issued by other governments are unlikely to displace the premier safe-haven status of US Treasuries for decades. But unless concerns about the resilience of markets for Treasuries are addressed, investors who anticipate a need to raise large amounts of cash quickly in a future crisis may reduce their preference for Treasuries. The equilibrium cost of funding the US government may correspondingly rise, and future financial crises may be worsened by the reduced ability of investors to obtain liquidity for the world's globally recognized safe financial asset, US Treasuries.

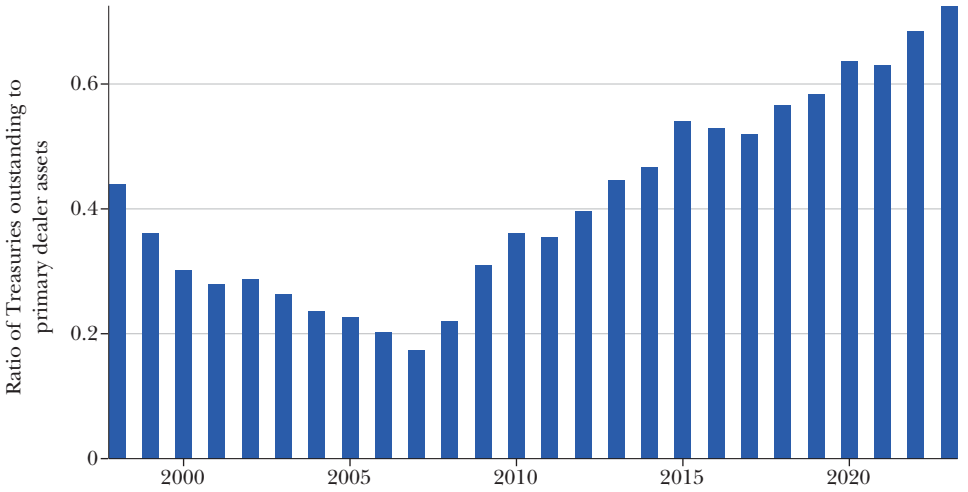
The remainder of this paper dives deeper into the design of the Treasury market, its intermediation bottlenecks, and relevant policy approaches. Whether Treasury markets will be functional in future crises, just when safe-haven investors most need liquidity, depends on improvements in market structure and regulation, as I will explain.

Background on the Functioning of the US Treasury Market

On an average day, over $1 trillion of Treasuries are traded by US dealers.[1] Among the hundreds of dealers in the US Treasury market, a small subset of "primary dealers," designated by the New York Fed, handle the lion's share of this trade. In early 2025, there were 25 primary dealers, the largest of which are run by

[1] This statement is based the TRACE data from the Financial Industry Regulatory Authority, January to March 2025, collected by Federal Reserve Bank of New York, and then accessed by SIFMA (the Securities Industry and Financial Markets Association).

Figure 1
The Ratio of US Treasury Securities Outstanding to Primary Dealer Assets, 1998–2023



Source: The data are from the Federal Reserve and public company filings. This figure updates Duffie (2023, Figure 2).
Note: The height of the bar for each year is the ratio of US Treasury securities outstanding to primary dealer assets. The sample period is 1998–2023. Assets are measured at the holding company level.

"globally systemically important banks."[2] Primary dealers form the core of the Treasury market. They commit to bid for large quantities in Treasury issuance auctions, and they intermediate the majority of trades of previously issued Treasury securities. The Federal Reserve Bank of New York also relies on the primary dealers when implementing its monetary policy operations in the Treasury market. Each day, primary dealers also provide trillions of dollars of secured lending to investors who provide their Treasuries as collateral.

Since the global financial crisis began in 2007, however, the total size of primary-dealer financial assets per dollar of outstanding Treasuries has shrunk by a factor of roughly four, as shown in Figure 1. This trend, representing shrinking relative market capacity, continues because of two underlying causes. First, the denominator of this ratio is dominated by assets of the globally systemically important banks, which are growing much more slowly because of stronger post-global financial crisis regulatory capital requirements. Second, larger US fiscal deficits have greatly expanded the numerator.

Robust capital regulations for globally systemically important banks are necessary for financial stability, but also reduce the incentives of dealers, acting

[2] For a current list of primary dealers, see the website of the New York Fed at https://www.newyorkfed.org/markets/primarydealers.html.

on behalf of their shareholders, to expand their balance sheets so as to accommodate customer demands for liquidity. A globally systemically important bank can only expand its balance sheet by issuing new equity, which usually reduces returns of existing shareholders, or by increasing its leverage through borrowing, which is limited by regulatory capital ratios (Duffie 2022). Borrowing to buy safe assets like Treasuries can also be costly to shareholders in the globally systemically important banks (Andersen, Duffie, and Song 2019). Further, in the post-crisis world, borrowing rates of these globally systemically important banks are higher relative to risk-free rates, because of a decline in the presumption by creditors that globally systemically important banks are too big to fail (Berndt, Duffie, and Zhu 2025). All of these factors have reduced the incentives for the primary dealers associated with globally systemically important banks to absorb investor demands for liquidity. As a consequence, during the COVID-19 crisis of March and April 2020, dealers were unable to meet the enormous investor demands to sell Treasuries (Bräuning and Stein 2024; Duffie et al. 2023; He, Nagel, and Song 2022).

Meanwhile, burgeoning federal budget deficits continue to expand the outstanding stock of Treasuries rapidly relative to US GDP. The Congressional Budget Office projects that by 2034, absent changes in legislation, the $17.2 trillion stock of Treasuries that was outstanding when COVID-19 struck in March 2020—already leaving primary dealers unable to provide the market with sufficient liquidity—will have grown to $42 trillion.[3]

In summary, the US Treasury market is intermediated mainly by primary dealers whose balance sheets have limited flexibility. The market's structure has become inadequate, given the huge growth in its size. When dealer balance sheets are filled to capacity by crisis selling, trading costs soar, allocative efficiency suffers, and investors who had purchased Treasuries in order to quickly and efficiently raise cash in a crisis are likely to be thwarted.

The Structure of the Treasury Market

In this section, I describe the structure of the US Treasury market, which is comprised of five parts: (1) the primary market, in which the US government finances its deficits by auctioning new Treasury securities; (2) the interdealer market, where dealers, interdealer brokers, and specialized high-frequency trading firms trade Treasuries with each other; (3) the customer-to-dealer trading market, where customers buy and sell Treasuries with dealers; (4) the repurchase ("repo") market, where investments in Treasuries are financed by collateralized borrowing; and (5) the Treasury futures market. Each of these market segments has its own

[3] These are measured as "federal debt held by the public," shown historically by the Federal Reserve Bank of St. Louis at https://fred.stlouisfed.org/series/FYGFDPUN and projected by the Congressional Budget Office at https://www.cbo.gov/publication/59710#:~:text=In%20CBO's%20projections%2C%20federal%20budget,reaches%20116%20percent%20of%20GDP.

submarkets and a rich set of institutional mechanisms that have evolved to handle large trade volumes at low cost. In the following section, I discuss regulatory and policy changes that could expand the intermediation capacity of the markets for US Treasuries.

Most of the fiscal deficits of the US government are met by issuing various kinds of US Treasury securities in the primary market. Faced with various types of investor clienteles having different respective objectives, the Treasury Department's Office of Debt Management has designed a wide menu of bills, notes, and bonds to offer to the market. Treasury "bills" with shorter maturities ranging from 4 to 17 weeks are auctioned weekly. One-year bills are auctioned monthly. New Treasury "notes" with maturities of two to seven years are auctioned monthly. New ten-year notes are issued quarterly, but there are monthly "reopening" auctions of the latest ten-year note. For example, the ten-year note issued in January is auctioned again in February and in March. A new ten-year note is then issued in April. These reopenings meet the preference by market participants for trading a large liquid stock of homogeneous notes. Initial offerings of Treasury bonds, with original maturities of 20 or 30 years, are auctioned quarterly. Acting as an agent for the Treasury, the Federal Reserve Bank of New York conducts these auctions and arranges for the settlement of cash payments from buyers to the Treasury's account at the Fed.

As the total Treasury debt has grown, the need to place so much debt has led the Treasury's Office of Debt Management to broaden the menu of Treasury securities offered to the market by introducing new maturity classes and new types of Treasuries. For example, 1998 saw the introduction of Treasury Inflation Protected Securities (TIPS) (Dudley, Roush, and Ezer 2009). By linking interest and principal payments to the Consumer Price Index, TIPS appeal to investors seeking a safe inflation hedge (Tobin 1963). TIPS are issued with original maturities of 5, 10, and 30 years. In 2013, the Treasury began issuing floating rate notes, whose successive interest payments vary over time with short-term market interest rates. Floating rate notes appeal to investors that prefer a medium-maturity Treasury security whose market value is relatively insensitive to changes in interest rates (Copic et al. 2014). Money market mutual funds are a natural clientele for floating rate notes, which are newly issued four times each year with original maturities of two years.

Immediately after being issued, a Treasury security is said to be "on the run." On-the-runs are extremely actively traded until a new security of the same maturity class is issued. Once that happens, the previous on-the-run security is said to be "off the run." Some on-the-runs are so popular in the secondary market that the same security is reissued in subsequent auctions in order to meet the high investor demand for liquid on-the-run benchmarks for each Treasury maturity class. Off-the-runs, being less frequently traded, have higher average trading costs than on-the-runs and, consequently, lower prices after adjustment for maturity differences and other effects (Duffie et al. 2023).

Treasuries are issued by auction to primary dealers and other direct bidders, which include foreign central banks and other large financial institutions. Given the massive increases in Treasury issuances and constraints on dealer balance sheets

associated with heightened capital regulations that were applied in the aftermath of the global financial crisis, the fraction of notes and bonds purchased at issuance by primary dealers has declined from more than 50 percent before the global financial crisis to merely 10 to 20 percent more recently (Wessel 2024a).

The large-menu approach, by which the Treasury auctions many types of securities, involves tradeoffs. A cost of this approach is the reduction in trading volumes of the many resulting off-the-run securities, implying higher secondary-market trading costs (Cochrane 2015). The benefit of a large menu of securities is the higher demand by each "clientele" of investors for securities that better meet their respective investment objectives (Vayanos and Vila 2021). For example, money-market mutual funds buy Treasury floating rate notes and bills. Managers of foreign-exchange reserves, by contrast, hold only about 7 percent of their Treasuries in the form of bills.[4] Insurance firms and pension funds prefer longer-maturity Treasuries as hedges against their long-term liabilities. Banks hold significant amounts of medium-term Treasury notes as investments that balance the overall risk and return of their liabilities and other assets (DeMarzo, Krishnamurthy, and Nagel 2024). Bond mutual funds often specialize in specific maturity sectors that appeal to their investors.

The inventories of Treasuries held by a dealer fluctuate in size with the dealer's auction purchases and customer trading demands. Dealers use various forms of interdealer trading to manage their resulting inventory imbalances and to profit by providing liquidity to each other (Fleming, Nguyen, and Rosenberg 2024). Dealers rely in part on electronic interdealer trade platforms such as BrokerTec, a limit-order-book market operated by CME Group that handles the majority of on-the-run trade.[5] The two main classes of participants on interdealer trade platforms are dealers and "principal trading firms," which use high-frequency trading algorithms and hold small inventories of Treasuries outside of trading hours. Given the burgeoning size of the Treasury market and the limited flexibility of the balance sheets of primary dealers associated with globally systemically important banks under the regulations enacted in the aftermath of the global financial crisis, the dominant class of participants on interdealer trade platforms has shifted from primary dealers to principal trading firms (Harkrader and Puglia 2020). The Securities and Exchange Commission (SEC) recently required some key principal trading firms to register as dealers. While this requirement will improve the transparency and stability of trading by the principal trading firms, the associated capital requirements may dampen their commitments of capital to the Treasury market, potentially harming market liquidity.[6]

[4] See the Treasury International Capital (TIC) dataset, accessible at https://ticdata.treasury.gov/resource-center/data-chart-center/tic/Documents/slt_table5.html.

[5] In addition to BrokerTec, CME Group owns various platforms for trading financial derivatives including the Chicago Mercantile Exchange (CME), the Chicago Board of Trade (CBOT), and the New York Mercantile Exchange (NYMEX).

[6] SEC regulations for the capital that primary dealers must hold do not recognize an increase in dealer capital unless the new capital is maintained for at least one year. This reduces the incentives of trading firms to provide downstream capital to their regulated dealer subsidiaries and thus impinges unnecessarily on market liquidity (Duffie 2024).

Large nondealer market participants, often called "the buyside," trade only with dealers and account for the majority of trade in the customer-to-dealer market. In effect, buyside firms do not have access to the interdealer market. The result is a core-periphery market structure in which dealers are completely connected with each other and buyside firms are connected only with dealers. This market structure has been sustained despite the improved competition and allocative efficiency that could be achieved with "all-to-all" trade venues, which would allow direct trade between Treasury market participants of all types (Chaboud et al. 2022; Allen and Witwer 2023). Households obtain their Treasuries mainly through asset-management vehicles such as mutual funds and hedge funds. Households can also buy new Treasury issuances directly from the Treasury, "noncompetitively," at the market-clearing Treasury auction prices.

The market for repurchase agreements ("repos") plays a key role in providing financing to bond-market investors across the US economy. Treasuries are the most popular class of bonds financed in the repo market. For example, a dealer could pledge Treasuries with a market value of $100 million to a money market mutual fund in exchange for $98 million of cash financing on an overnight repo. The corresponding 2 percent "haircut" protects the money-market fund from the risk of default by the dealer. The next day, the dealer pays back $98 million in cash plus one day of interest, and the money fund returns the Treasury securities collateral to the dealer. Many repos have no haircuts at all (Hempel et al. 2023)! This reflects the relative safety of Treasury collateral and the fact that some investors have access to competing sources of dealer repo financing and are able to exploit competition among dealers to negotiate no-haircut deals.

The repo market involves a large volume of "cash-futures basis" trading, a form of arbitrage conducted by hedge funds. When considering whether to enter a basis trade, a hedge fund compares the current Treasury futures prices at the Chicago Board of Trade (CBOT) to the expected cost of purchasing Treasury securities and financing them in the repo market until the futures contract matures. The difference between these alternatives is called "the basis." If the basis is large enough, as has often been the case in recent years, the hedge fund can arbitrage by shorting Treasury futures, buying Treasury securities, financing the Treasuries in the repo market, and delivering the Treasuries to meet the delivery obligations of the short futures position. The basis trade position can also be liquidated before the futures delivery date. There is a risk that between the inception of a basis trade and its delivery date, the basis may get even larger and force the hedge fund to provide additional margin. In sufficiently severe scenarios, some hedge funds could be unable to meet the call for additional margin and fail, potentially causing a large fire-sale of basis-trade positions. The extent of basis trades is difficult to measure but has been large enough to raise financial-stability concerns (Barth, Kahn, and Mann 2023). Earlier this year, CME Group (2025) announced the launch of a platform for trading combinations of Treasury securities and Treasury futures contracts, thus accommodating the ability to conduct a basis trade with a single trade execution.

Interest-rate derivatives are widely used for managing the risks of positions in Treasuries and other bonds, or as a substitute for purchasing Treasuries. For example, a combination of Treasury futures contracts and interest rate swaps (derivatives that contract for the exchange of a stream of fixed interest payments for a stream of floating interest payments) is often used as a substitute for purchasing Treasury securities and financing them in the repo market. A mixture of accounting rules and capital regulations discourage mutual funds, pension funds, and insurance firms from financing Treasury purchases in the repo market. These large institutional investors thus often prefer to obtain their desired exposure to interest rates through derivatives, which receive more beneficial accounting or capital treatment. This avoidance of direct purchases of Treasuries by many large investors has elevated long-term Treasury yields significantly relative to the yields implied by interest-rate derivatives. These differences in yields, called "swap spreads," now exceed 50 basis points (0.5 percent) at maturities of ten years or more. Such large swap spreads represent a significant frictional upward distortion in Treasury yields, thus a substantial elevation in the cost to taxpayers for financing US fiscal deficits.

Large swap spreads also elevate the incentive for hedge funds to conduct basis trading. If enough capital is allocated to basis-trade arbitrage, the yields of Treasuries would fall toward the yields implied by derivatives prices, potentially saving US taxpayers a large amount of interest expense on Treasury debt. Currently, each basis point by which Treasury yields are lowered through basis-trade arbitrage reduces US government interest expense by roughly $3 billion annually, a potential saving that will rise proportionately with the growth of the Treasury market.[7] Regulators therefore face the challenge of ensuring that basis trading is not an excessive risk to financial stability while at the same time "hoping" that there is enough basis-trade arbitrage to bring Treasuries prices up into line with the prices implied by derivatives.

Concerns over the Resilience of Markets for Treasury Securities

How can the safe-haven status of US Treasuries and its related social benefits be safeguarded? Sufficiently reducing US fiscal deficits so that the existing primary dealers can intermediate the market seems like wishful thinking. Congress has shown little restraint here. But some progress has been made with redesigning the Treasury market so that its liquidity is more resilient to market stresses. After the COVID-19 shock of March 2020 revealed alarming Treasury market dysfunction, the US official sector put several types of reforms into motion (Interagency Working Group 2023). These reforms include (1) broadening the requirement

[7] I am grateful to Sam Wycherley for this estimate, from his research in progress. Net US Treasury interest expense was $881 billion in 2024 and is projected to reach $952 billion in 2025. Expressing net interest expense as an annual interest rate, the Treasury paid 3.12 percent on its debt in 2024 and is projected to pay 3.16 percent in 2025.

to clear Treasuries trades centrally; (2) improving post-trade price transparency; (3) increasing access to central-bank liquidity for Treasury securities under stressed-conditions; and (4) conducting Treasury "buybacks." Buybacks are auctions in which the Treasury purchases older less-liquidly traded off-the-run Treasuries. These purchases are funded by issuing additional amounts of new liquid on-the-run Treasuries. In this section, I first focus on these reforms. I then turn to two other potentially important additional improvements in Treasury markets that are still under discussion: (1) a reduced role for leverage-based regulatory capital requirements, which in their current form ignore the riskiness of dealer assets when determining bank capital buffers and therefore discourage the intermediation of safe assets like Treasury securities, and (2) the emergence of all-to-all trade in Treasury securities.

Central Clearing

A Treasuries transaction, whether an outright trade or a repurchase agreement, is said to be "settled" when the promised cash and securities are actually delivered. Currently, the majority of Treasury-market settlements are conducted directly between the original two trade counterparties—for example, between two dealers, or between a dealer and a customer. By 2027, a new Securities and Exchange Commission rule will require a wide swath of Treasuries transactions to be settled by a clearinghouse (also known as a "central counterparty"), which acts as the seller to the original buyer and as the buyer to the original seller.[8] The clearinghouse offers a guarantee: if one of the original counterparties fails to perform at settlement, then the clearinghouse will complete the settlement. The members of the clearinghouse, mostly large dealers, collectively absorb any resulting losses that remain after applying the margin supplied to the clearinghouse by the defaulting counterparty.

Currently, the Fixed Income Clearing House is the only regulated US clearinghouse for Treasuries transactions. However, the prospect of an enormous increase in the volume of central clearing, due to the growth of the Treasury market and the Securities and Exchange Commission's new central clearing mandate, has incited the CME Group, Intercontinental Exchange (ICE), and London Clearing House (LCH) all to announce plans for their potential introduction of clearinghouses for Treasuries.

Central clearing makes safer and more efficient use of dealer balance-sheet space. By central clearing their transactions, dealers can reduce their total financial commitments to settle their trades by offsetting purchases from some counterparties with sales to others. For example, with central clearing, a dealer who purchases $200 million of Treasuries from a mutual fund and sells $150 million of the same securities to an insurance firm is left with only a $50 million commitment to the clearinghouse. Without central clearing, the dealer's total settlement exposure—that is, the total risk that counterparties will not meet their commitments to deliver

[8] In February, the SEC (2025) announced a delay in the completion date from June 2026 to June 2027.

cash or securities—is $350 million. Fleming and Keane (2021) estimate that central clearing could generate large reductions in dealer settlement exposures, even up to 70 percent smaller exposures in extreme conditions.

By contract, a counterparty's failure to meeting its settlement commitment on time is often cured with a delayed settlement rather than a default. However, a failure by A to deliver to B on time can cause B to fail to C, then C to fail to D, and so on. These "daisy chains" of bilateral failures, which increased by hundreds of billions of dollars during the COVID-19 shock of March 2020, are significantly reduced by central clearing (Duffie 2020; Fleming and Keane 2021). Without central clearing, daisy-chain failures can propagate unpredictably through a complex web of principal trading firms, trade platform operators, interdealer brokers, dealers, and buyside firms, leading to significant additional risk that counterparties will end up defaulting, potentially destabilizing Treasuries markets.

Wider use of central clearing should lead to an overall increase in trust of market resilience for Treasury securities by increasing market transparency, reducing counterparty risk, and reducing dangerous failures to deliver securities promised on a trade. Through the netting of original purchases and sales at a clearinghouse, central clearing also reduces regulatory capital requirements for dealers. By a quirk of regulatory accounting, however, commitments to settle Treasury securities purchases and sales do not count toward capital requirements, whereas economically equivalent exposures to repo settlements do count toward capital requirements. Because of this arcane inconsistency, the Securities and Exchange Commission's new central clearing rule will not actually lead to much incremental reduction in leverage-based capital requirements (Bowman, Huh, and Infante 2024). Increased use of central clearing also implies a greater adverse impact on the economy if a clearinghouse were to fail. The largest US central counterparties, including the Fixed Income Clearing House, have been designated by the Financial Stability Oversight Council as systemically important, and are effectively "too big to fail" (Yadav and Younger 2025).

At present, the Federal Reserve is not a member of any clearinghouse for US Treasuries. However, there is a strong financial-stability rationale for the Fed to link itself with any systemically important Treasuries clearinghouse. This would benefit the resilience of the US Treasury market, and financial stability more generally. Although offering emergency central-bank liquidity to a clearinghouse presents a moral hazard, focusing the Fed's liquidity provision to the case of US Treasuries significantly lowers this moral hazard. If a clearinghouse needs more cash financing for Treasuries than it can obtain in a crisis from its private-sector clearing members, it would be fully in keeping with the traditional lender-of-last-resort role of a central bank to take Treasuries as collateral against this lending of last resort. Moreover, if the Fed is a member of a Treasuries clearinghouse, the Fed would be positioned to offer emergency liquidity to a wider set of market participants without the need to manage the risk of its trade settlements with each individual firm. There should be no requirement for the Fed to be a loss-sharing member of the clearinghouse, given that the Fed itself is never at risk of insolvency.

As an additional benefit of Fed participation in the central clearing of US Treasuries, primary dealers could take advantage of using central clearing to net their trades with the Fed against their trades with other counterparties. This would expand the effective capacity of primary dealers to provide liquidity to Treasury market participants, especially in a crisis that induces large customer demands to sell Treasuries to primary dealers and that leads to large purchases of Treasuries by the Fed from the same dealers. The COVID-19 crisis of March–April 2020 caused just this scenario, but primary dealers were unable to use central clearing to net their customer trades against their Fed trades because the Fed does participate in central clearing. This is a significant missed opportunity.

Post-Trade Price Transparency

Regulators are slowly moving toward a plan for improving post-trade price transparency in the market for US Treasury securities by publishing trade prices and quantities shortly after each trade (Liang 2022). Post-trade price transparency will likely improve competition and allocative efficiency. With the ability to compare the prices offered by dealers with recently reported transaction prices, buyside firms will be in a better position to shop for price improvement with other dealers or on a trade platform (Duffie, Dworczak, and Zhu 2017). The efficiency with which dealers are matched to trades will improve, likely expanding the intermediation capacity of the market. Eventually, greater post-trade price transparency will also speed up the emergence of all-to-all trade. Faced with greater price transparency, some dealers will find it more effective to compete for revenues by helping their customers execute some of their trades on competitive platforms, rather than relying only on trading bilaterally with their customers.

Transaction-level data in the market for Treasury securities are already reported to regulators in the Trade Reporting and Compliance Engine (TRACE) run by the Financial Industry Regulatory Authority (FINRA), which is a self-regulatory organization for broker-dealer firms that plays a role in regulating the securities industry. However, most of these data are not yet being publicly released, not even with a delay (Brain et al. 2018). Dealers have expressed the concern that if a large trade is published immediately, some market participants may attempt to profit by predicting follow-on trades of the same type and trade ahead of the dealers, a form of "front running" that reduces the incentive of dealers to offer attractive prices to their original customers. To mitigate potential cost of front-running, the sizes of larger trades could be censored from public reporting by TRACE, a practice that has been applied effectively in corporate-bond transaction reporting by TRACE for the past two decades.

At present, dealers in Treasury securities benefit from the absence of post-trade price transaction transparency, through reduced competition. The fact that transactions have not been reported by the Trade Reporting and Compliance Engine, even months after the trade, implies that dealers have been protected not only from front-running, but also from the ability of their customers to later analyze their trade execution costs—that is, to determine whether the prices they achieved

from their dealer on past trades were close to those achieved elsewhere in the market at the same time. Shielding dealers from this form of delayed scrutiny is a strange choice by regulators. Perhaps the government is aiming to protect the profit margins of primary dealers so that they will continue to commit capital to Treasury market. However, protecting dealers from competition will not support the structural market improvements that are needed to accommodate ever-larger peaks in market volumes.

Official-Sector Backstops

At the 2021 US Treasury Market Conference, John Williams (2021), President of the Federal Bank of New York, drew on the experiences of March 2020 by observing: "[T]wo lessons are clear. First, the unforeseeable and unpredictable will happen, and can result in significant stresses in the Treasury and related markets that may spread to broader financial conditions. Second, when disruptions have been sufficiently severe and persistent, the market has not been able to quickly self-correct without official-sector intervention."

Beyond purchasing trillions of dollars worth of Treasuries to restore market functioning after the COVID-19 shock of 2020, the Federal Reserve introduced two new facilities that provide repo financing for Treasuries in times of financial stress. The Standing Repo Facility (SRF) targets this support to primary dealers and banks. The Foreign and International Monetary Authorities (FIMA) Repo Facility allows foreign monetary authorities with a custodial account at the Federal Reserve Bank of New York to obtain financing for the Treasuries held in their custodial accounts.[9]

The Group of Thirty (2021) and Hubbard et al. (2021) suggest that a broader set of market participants should have access to the Standing Repo Facility. In 2024, Bank of England (2024) instituted a contingent financing facility for UK government bonds that is available to a wider set of non-bank financial institutions.

In a deep crisis, central bank financing of Treasury securities is not necessarily sufficient. In some crises, including the COVID-19 shock of March 2020, central banks also stepped in with direct emergency purchases of government securities to restore market functioning (Duffie 2020; Garbade 2021). Rather than reacting during a crisis on an ad-hoc basis, the Fed could set up a transparent facility that stands ready, as a last resort, to purchase government securities in future crises (Duffie and Keane 2023). With knowledge of the existence and structure of such a backstop to market liquidity, investors would be willing to pay more for Treasuries when they are auctioned, reducing the cost to taxpayers of financing the US government. In a crisis, investors would be less likely to race each other to get liquidity from limited dealer balance sheets.

[9] In the Silicon Valley Bank crisis of 2023, the Bank Term Funding Program provided financing for Treasuries held by banks (see https://www.federalreserve.gov/financial-stability/bank-term-funding-program.htm). For a discussion of that episode in this journal, see Metrick (2024). For FAQs about the Standing Repo Facility, see https://www.newyorkfed.org/markets/repo-agreement-ops-faq. For FAQs about FIMA, see https://www.federalreserve.gov/monetarypolicy/fima-repo-facility-faqs.htm.

A potential challenge with Fed emergency purchases of Treasuries to support market functioning is that this can sometimes act at cross purposes with monetary policy. In principle, fiscal purchases directly by government finance ministries and Treasury departments are an alternative to the conduct of monetary policy through market-function purchases and corresponding movements of the target interest rate (Duffie and Keane 2023). In some instances, purchases of government securities by a central bank to restore market functioning are not naturally aligned with optimal monetary policy and can make policy-related communications more challenging. On September 22, 2022, for instance, the Monetary Policy Committee of the Bank of England voted to begin selling gilts (the British equivalent of Treasury securities) for the purpose of quantitative tightening. But within a day of this announcement, a UK fiscal policy shock led to sharp jumps in UK yields and fire sales of gilts by "liability-driven investors," comprised largely of pension funds that put a high priority on matching their investment portfolios to their future pension liabilities. Large sales of gilts by these liability-driven investors destabilized the market, leading the Financial Policy Committee of the Bank of England to institute a program of gilt purchases that soon restored market stability (Hauser 2022). As Bank of England Governor Andrew Bailey (2022) has stated, "There may appear to be a tension here between tightening monetary policy as we must, including so-called Quantitative Tightening, and buying government debt to ease a critical threat to financial stability. This explains why we have been clear that our interventions are strictly temporary and have been designed to do the minimum necessary."

Buybacks

The US Department of the Treasury recently instituted a Treasury securities "buyback" program (Frost 2024), under which the Treasury uses auctions to purchase less liquid off-the-run securities from primary dealers. Although the stated primary purposes of this program do not include crisis buying to improve the functionality of the Treasury market, a related benefit is that buybacks reduce the stock of illiquid off-the-run Treasuries that would be sold to dealers during a dash for cash. Because off-the-runs are less frequently traded, they are more difficult for dealers to resell and are therefore more likely than on-the-run Treasuries to "clog" dealer balance sheets in a crisis. The Treasury Department's buyback program therefore reduces the extent to which dealer balance sheets could become clogged with illiquid securities in a crisis. Initially, however, the size of the buyback program is modest.

Reducing the Distortionary Effect of Leverage-Based Capital Requirements

A pure leverage-based capital rule requires that a bank's capital exceeds a fixed fraction of its total assets—without consideration of the riskiness of the assets. Relative to foreign versions of leverage-based capital rules, the US enhanced Supplementary Leverage Ratio leans particularly heavily on the largest bank holding companies, and thus on the largest primary dealers in the US Treasury market (Tarullo 2023). Leverage-based capital rules are distortionary. For instance, the US rule requires significant capital to cover "losses" even on Federal Reserve deposits (that is, bank

reserves held at the Fed), even though reserves have literally no risk of loss and are instantly liquid! As another example, a Treasury repo is nearly risk-free, but requires the same capital buffer per dollar of assets as a risky real-estate loan or any other asset. A reduced role for leverage-based capital regulation, or an exemption from the Supplementary Leverage Ratio for at least Treasuries and reserves, would allow more efficient use of dealer balance sheets and improve the incentives of dealers to absorb more customer Treasury positions in a crisis (Duffie 2022; Group of Thirty 2021).

In March 2020, amidst the crisis selling of Treasury securities, the balance sheets of the largest dealers became heavily loaded with Treasury securities, repos, and reserves. The Supplementary Leverage Ratio thus required primary dealers at globally systemically important banks to allocate a large amount of their capital to their positions in Treasuries, Treasury repos, and reserves, even though these are on average very safe assets. In this way, the Supplementary Leverage Ratio inefficiently reduced the incentives of dealers to absorb additional investor positions. On April 1, 2020, regulators recognized this impediment to market liquidity, exempting reserves and Treasuries from the Supplementary Leverage Ratio. Those dealers at globally systemically important banks that had been relatively more constrained by the Supplementary Leverage Ratio quickly reacted by committing relatively more of their balance sheets to Treasury positions (Bräuning and Stein 2024). However, this exemption lasted for only one year.

An exemption from the Supplementary Leverage Ratio for Treasuries (and perhaps reserves) was supported in early 2025 by the Chair of the Federal Reserve (as reported by Grossman 2025) and by the US Treasury Secretary (Bessent 2025). The reduction in dealer capital associated with such an exemption should ideally be offset by increasing risk-based capital requirements. Without such an offset, dealer capital levels could decline, leading dealer borrowing costs to rise correspondingly, eventually increasing the costs to dealers for financing their inventories through higher credit spreads. This would reduce the incentives of dealers to offer liquidity to financial markets (Andersen, Duffie, and Song 2019), although the net effect of an exemption from the Supplementary Leverage Ratio on Treasury market liquidity would likely still be positive. Today, the best-capitalized dealers have lower funding costs and lower required returns on equity than less well-capitalized dealers. Better capitalized dealers can therefore more efficiently provide liquidity to their customers.

All-to-All Trade

Eventually, all-to-all platforms for trading Treasury securities will become available, so that most Treasury market participants will be able to trade with most other market participants (Chaboud et al. 2022). Dealer balance sheets will no longer be required to absorb all buyside trades, thus expanding market intermediation capacity. While investors will continue to conduct many of their trades directly with dealers, the option to trade on all-to-all trade platforms will improve competition and matching efficiency (Allen and Wittwer 2023; Kutai, Nathan, and Wittwer 2025).

Dealers have shown no interest in supporting all-to-all trade, given the implied increase in competition they would face from all-to-all trade venues. It is reasonable to predict, however, that all-to-all trade will eventually incite such a large expansion in trade volumes that total dealer profitability would likely rise, despite the resulting reduction of dealer profit on the average trade. For example, prior to the 1973 introduction of exchange trading of equity options on the Chicago Board Options Exchange (CBOE), dealers handled virtually all trade in equity options. The first month of CBOE had *more trade volume than any prior year of dealer-only trade* (Duffie 2019). Exchange-traded equity options volumes now exceed 1973 levels by a multiple of approximately 10,000. Dealers have benefited greatly from the expansion of trade volumes in option markets caused by all-to-all trade, even though the associated increase in market competition has caused a reduction of dealer profit margins on each trade.

All-to-all trade can be implemented in various ways. For example, an electronic trade platform like BrokerTec (which is owned by the CME Group) could encourage buyside firms to participate in its limit-order-book market, following the form of all-to-all trade used on futures and stock exchanges. New trade platforms could also offer all participants the ability to post requests for price quotes from all other platform participants, an approach that is becoming more popular in the corporate bond market (Hendershott, Livdan, and Schürhoff 2021). Alternatively, trade platforms could organize occasional double auctions for a wide set of market participants.

All-to-all trade is likely to be promoted not only by post-trade price transparency, but also by central clearing. Without an official clearinghouse to handle trade settlement, all-to-all trade platform operators end up serving as a de facto clearinghouse, but their cost for committing sufficient capital to settle large volumes of trade safely would make this approach prohibitively costly.

Currently, there is a "done-with" norm in the US Treasury market, meaning that an investor who arranges with a dealer for the central clearing of a trade must also conduct the trade with that same dealer. This done-with approach is controversial because it reduces trade competition (Clancy 2025). Done-with trade also lowers incentives to conduct all-to-all trade. Anonymous central clearing and greater flexibility for "done-away" trades would promote all-to-all trade and thus increase market capacity.

Concluding Remarks

US Treasuries will likely remain the world's premier safe-haven investment for decades to come. No serious contender is in sight. The key questions are how effective these safe-haven services will be and how much more costly it will be for the US government to finance itself if it does not sufficiently shore up the resilience of the Treasury market. While planned improvements in regulation and market design will make more effective use of dealer balance sheets, this will not be enough. The market for Treasury securities is simply growing too large to rely exclusively

on dealers to intermediate investor trades. The decision by the Securities and Exchange Commission to extend the scope of required central clearing was a major step forward. Large improvements in price transparency and market structure are also needed. Regulations should promote (although not mandate) all-to-all trade. The role of capital requirements based purely on leverage ratios should be reduced relative to risk-based capital requirements. For confidence in the safe-haven services of US Treasuries, the intermediation capabilities of the market will need to be more resilient to crisis selling, especially as the potential size of crisis selling expands dramatically with the overall rise in the quantity of Treasury securities.

Even with the suggested improvements in market design and regulation, there will always be potential crises severe enough to cause dysfunction in the Treasury market. The ability of the Federal Reserve to backstop the market with lending of last resort and Treasury purchases should go beyond the traditional perimeter of banks and primary dealers. In particular, given the huge volume of trades that will be settled by clearinghouses, the Fed should prepare to settle trades at a clearinghouse and to provide liquidity for Treasuries to a wider set of market participants and to systemically important clearinghouses.

■ *I am grateful for collaboration on this subject with Rafael Berriel Abreu, Miguel Chumbo, Michael Fleming, Tim Geithner, Frank Keane, Claire Nelson, Pat Parkinson, Or Shachar, Jeremy Stein, Peter Van Tassel, and Sam Wycherly, and for related discussions over the years with many others including Larry Bernstein, Josh Frost, Ken Garbade, Linda Goldberg, Jason Granet, Sam Hanson, Graham Harper, Frank Keane, Don Kohn, Nellie Liang, Lori Logan, Jun Pan, Tom Pluta, Myron Scholes, Hyun Shin, Andreas Schrimpf, Adi Sundaram, Colin Teichholtz, Jessica Wachter, David Wessel, Don Wilson, Nate Wuerffel, and Haoxiang Zhu. I am also grateful for assistance on related research from Austin Bennett, Rania Ezzane, Manan Gupta, Renhao Jiang, Isabel Krogh, Taimin Liao, Allen Liu, Hala Moussawi, and Rany Stephan. I am especially thankful for extensive suggestions by JEP editors Jonathan Parker, Nina Pavcnik, and Timothy Taylor. All opinions expressed are my own. I am an independent director of Dimensional US Mutual Funds Board.*

References

Allen, Jason and Milena Wittwer. 2023. "Centralizing Over-the-Counter Markets?" *Journal of Political Economy* 131 (12): 3310–51.

Andersen, Leif, Darrell Duffie, and Yang Song. 2019. "Funding Value Adjustments." *Journal of Finance* 74 (1): 145–92.

Arslanalp, Serkan, Barry Eichengreen, and Chima Simpson-Bell. 2022. "The Stealth Erosion of Dollar Dominance and the Rise of Nontraditional Reserve Currencies." *Journal of International Economics* 138: 103656.

Bailey, Andrew. 2022. "Monetary Policy and Financial Stability Interventions in Difficult Times." Speech, G30 37th Annual International Banking Seminar, Washington, DC, October 15. https://www.bankofengland.co.uk/speech/2022/october/andrew-bailey-opening-remarks-and-panellist-37th-annual-international-banking-seminar.

Bank of England. 2024. "Contingent NBFI Repo Facility (CNRF)." Bank of England Explanatory Note, July 24. https://www.bankofengland.co.uk/markets/market-notices/2024/july/contingent-nbfi-repo-facility-explanatory-note.

Barone, Jordan, Alain Chaboud, Adam Copeland, Cullen Kavoussi, Frank Keane, and Seth Searls. 2022. "The Global Dash for Cash: Why Sovereign Bond Market Functioning Varied across Jurisdictions in March 2020." Federal Reserve Bank of New York Staff Report 1010.

Barth, Daniel, R. Jay Kahn, and Robert Mann. 2023. "Recent Developments in Hedge Funds' Treasury Futures and Repo Positions: Is the Basis Trade 'Back'?" FEDS Notes, August 30. https://www.federalreserve.gov/econres/notes/feds-notes/recent-developments-in-hedge-funds-treasury-futures-and-repo-positions-20230830.html?mod=livecoverage_web.

Bernardini, Marco, and Annalisa De Nicola. 2025. "The Market Stabilization Role of Central Bank Asset Purchases: High-Frequency Evidence from the COVID-19 Crisis." *Journal of International Money and Finance* 152: 103257.

Berndt, Antje, Darrell Duffie, and Yichao Zhu. 2025. "The Decline of Too Big to Fail." *American Economic Review* 115 (3): 945–74.

Bessent, Scott. 2025. "Remarks at the Economic Club of New York." Economic Club of New York, March 6. https://home.treasury.gov/news/press-releases/sb0045.

Bowman, David, Yesol Huh, and Sebastian Infante. 2024. "Balance-Sheet Netting in US Treasury Markets and Central Clearing." Finance and Economics Discussion Series Working Paper 2024-057.

Brain, Doug, Michiel De Pooter, Dobrislav Dobrev, Michael Fleming, Pete Johansson, Colin Jones, Frank Keane, Michael Puglia, Liza Reiderman, Tony Rodrigues, and Or Shachar. 2018. "Unlocking the Treasury Market through TRACE." FEDS Notes, September 28. https://www.federalreserve.gov/econres/notes/feds-notes/unlocking-the-treasury-market-through-trace-20180928.html.

Bräuning, Falk, and Hillary Stein. 2024. "The Effect of Primary Dealer Constraints on Intermediation in the Treasury Market." Federal Reserve Bank of Boston Working Paper 24-7.

Buiter, Willem H., Stephen G. Cecchetti, Kathryn M. Dominguez, and Antonio Sánchez Serrano. 2023. "Stabilising Financial Markets: Lending and Market Making as a Last Resort." European Systemic Board Report of the Advisory Scientific Committee 13.

Chaboud, Alain, Ellen Correia-Golay, Caren Cox, Michael J. Fleming, Yesol Huh, Frank M. Keane, Kyle Lee, Krista Schwarz, Clara Vega, and Carolyn Windover. 2022. "All-to-All Trading in the US Treasury Market." Federal Reserve Bank of New York Staff Report 1036.

Clancy, Luke. 2025. "PTFs Clash with Banks over 'Done-Away' US Treasury Clearing." Risk.net, March 14. https://www.risk.net/risk-management/7961212/ptfs-clash-with-banks-over-done-away-us-treasury-clearing.

CME Group. 2025. "Press Release: CME Group to Launch BrokerTec U.S. Treasury Central Limit Order Book in Chicago to Streamline Trading between Cash and Futures Markets." CME Group, March 20. https://www.cmegroup.com/media-room/press-releases/2025/3/20/cme_group_to_launchbrokertecustreasurycentrallimitorderbookinchi.html.

Cochrane, John. 2015. "A New Structure for US Federal Debt." In *The $13 Trillion Question: How America Manages Its Debt*, edited by David Wessel, 91–146. Brookings Institution Press.

Copic, Ezechiel, Luis Gonzalez, Caitlin Gorback, Blake Gwinn, and Ernst Schaumburg. 2014. "Introduction to the Floating-Rate Note Treasury Security." *Liberty Street Economics* (blog), April 21. https://libertystreeteconomics.newyorkfed.org/2014/04/introduction-to-the-floating-rate-note-treasury-security/.

DTCC. 2024. *The US Treasury Clearing Mandate: An Industry Pulse Check*. Depository Trust and Clearing Corporation

Davis, J. Scott. 2023. "Treasuries' Allure as Safe Haven Noted in Short Maturities, Not in Long Bonds." Federal Reserve Bank of Dallas (blog), June 27. https://www.dallasfed.org/research/economics/2023/0627.

DeMarzo, Peter M., Arvind Krishnamurthy, and Stefan Nagel. 2024. "Interest Rate Risk in Banking." NBER Working Paper 33308.

Diamond, William, and Peter Van Tassel. 2024. "Risk-Free Rates and Convenience Yields around the World." http://dx.doi.org/10.2139/ssrn.4048083.

Du, Wenxin, Joanne Im, and Jesse Schreger. 2018. "The US Treasury Premium." *Journal of International Economics* 112: 167–81.

Dudley, William C., Jennifer Roush, and Michelle Steinberg Ezer. 2009. "The Case for TIPS: An Examination of the Costs and Benefits." *FRBNY Economic Policy Review* 15 (1): 1–17.

Duffie, Darrell. 2015. "Discussion of 'A New Structure for US Federal Debt,' by John Cochrane." In *The $13 Trillion Question: How America Manages Its Debt*, edited by David Wessel, 139–46. Brookings Institution Press.

Duffie, Darrell. 2019. "Report in Support of Class Plaintiffs' Motion for Class Certification". In *re: Interest Rate Swaps Antitrust Litigation*, 16-MD-2704 (S.D.N.Y.), originally filed under seal on February 20, 2019, and filed in redacted form on March 7, 2019 (Dkt. No. 725–2). https://www.docketbird.com/court-documents/In-re-Interest-Rate-Swaps-Antitrust-Litigation/Exhibit-Expert-Report-of-Professor-Darrell-Duffie/nysd-1:2016-md-02704-00725-002.

Duffie, Darrell. 2020. "Still the World's Safe Haven? Redesigning the US Treasury Market after the COVID-19 Crisis." Brookings Institution Hutchins Center Working Paper 62.

Duffie, Darrell. 2022. *Fragmenting Markets: Post-Crisis Bank Regulations and Financial Market Liquidity*. De Gruyter.

Duffie, Darrell. 2023. "Resilience Redux in the US Treasury Market." In *Structural Shifts in the Global Economy, A Symposium Sponsored by Federal Reserve Bank of Kansas, Jackson Hole, Wyoming, August 24–26, 2023*, 77–119. Federal Reserve Bank of Kansas City.

Duffie, Darrell. 2024. "Comment to the Securities and Exchange Commission on Its Proposed Rulemaking on the Definition of Dealer and Government Securities Dealer." File No. S7-12-22. January 10. https://www.sec.gov/comments/s7-12-22/s71222-371039-899222.pdf.

Duffie, Darrell, Piotr Dworczak, and Haoxiang Zhu. 2017. "Benchmarks in Search Markets." *Journal of Finance* 72 (5): 1983–2044.

Duffie, Darrell, Michael Fleming, Frank Keane, Claire Nelson, Or Shachar, and Peter Van Tassel. 2023. "Dealer Capacity and US Treasury Market Functionality." Federal Reserve Bank of New York Staff Report 1070.

Duffie, Darrell, and Frank Keane. 2023. "Market-Function Asset Purchases." Federal Reserve Bank of New York Staff Report 1054.

Federal Reserve Board. 2020. *Financial Stability Report*. Federal Reserve Board.

Fleming, Michael J. 2021. *The Netting Efficiencies of Marketwide Central Clearing*. Federal Reserve Bank of New York.

Fleming, Michael J., Haoyang Liu, Rich Podjasek, and Jake Schurmeier. 2022. "The Federal Reserve's Market Functioning Purchases." *FRBNY Economic Policy Review* 28 (1): 210–41.

Fleming, Michael, Giang Nguyen, and Joshua Rosenberg. 2024. "How Do Treasury Dealers Manage Their Positions?" *Journal of Financial Economics* 158: 103885.

Frost, Joshua. 2024. "Remarks by Assistant Secretary for Financial Markets Joshua Frost on Recent Progress by the Inter-Agency Working Group on Treasury Market Surveillance." Speech, Federal Reserve Bank of New York's Annual Primary Dealer Meeting, US Department of the Treasury, May 8. https://home.treasury.gov/news/press-releases/jy2328.

Garbade, Kenneth D. 2021. *After the Accord: A History of Federal Reserve Open Market Operations, the US Government Securities Market, and Treasury Debt Management from 1951 to 1979*. Cambridge University Press.

Goldberg, Linda S., and Oliver Hannaoui. 2024. "Drivers of Dollar Share in Foreign Exchange Reserves." Federal Reserve Bank of New York Staff Report 1087.

Goldberg, Linda, and Fabiola Ravazzolo. 2021. "Do the Fed's International Dollar Liquidity Facilities Affect Offshore Dollar Funding Markets and Credit?" *Liberty Street Economics* (blog), December 20. https://libertystreeteconomics.newyorkfed.org/2021/12/do-the-feds-international-dollar-liquidity-facilities-affect-offshore-dollar-funding-markets-and-credit/.

Group of Thirty's Working Group on Treasury Market Liquidity. 2021. *US Treasury Markets: Steps toward Increased Resilience*. Group of Thirty.

Grossman, Matt. 2025. "Powell Says Easing Bank Capital Requirements Could Help Treasury Market." *Wall Street Journal*, February 12. https://www.wsj.com/livecoverage/cpi-report-today-inflation-stock-market-02-12-2025/card/powell-says-easing-bank-capital-requirements-could-help-treasury-market-s2FgvVH2xlHnNChGEb3o.

Harkrader, James Collin, and Michael Puglia. 2020. "Principal Trading Firm Activity in Treasury Cash Markets." FEDS Notes, August 4. https://www.federalreserve.gov/econres/notes/feds-notes/principal-trading-firm-activity-in-treasury-cash-markets-20200804.html.

Hauser, Andrew. 2022. "Thirteen Days in October: How Central Bank Balance Sheets Can Support Monetary and Financial Stability." Speech, ECB's 2022 Conference on Money Markets, Bank of England, November 4. https://www.bankofengland.co.uk/speech/2022/november/andrew-hauser-keynote-speech-at-the-european-central-bank-conference-on-money-markets.

He, Zhiguo, and Arvind Krishnamurthy. 2020. "Are US Treasury Bonds Still a Safe Haven?" In *NBER Reporter*, Iss. 3, 20–24. NBER.

He, Zhiguo, Stefan Nagel, and Zhaogang Song. 2022. "Treasury Inconvenience Yields during the COVID-19 Crisis." *Journal of Financial Economics* 143 (1): 57–79.

Hempel, Samuel J., R. Jay Kahn, Robert Mann, and Mark E. Paddrik. 2023. "Why Is So Much Repo Not Centrally Cleared? Lessons from a Pilot Survey of Non-centrally Cleared Repo Data." Office of Financial Research Brief 23-01.

Hendershott, Terrence, Dmitry Livdan, and Norman Schürhoff. 2021. "All-to-All Liquidity in Corporate Bonds." Swiss Finance Institute Research Paper 21-43.

Hubbard, Glenn, Donald Kohn, Laurie Goodman, Kathryn Judge, Anil Kashyap, Ralph Koijen, Blythe Masters, Sandie O'Connor, and Kara Stein. 2021. *Task Force on Financial Stability*. Hutchins Center on Fiscal and Monetary Policy at Brookings.

Interagency Working Group. 2023. *Enhancing the Resilience of the US Treasury Market: 2023 Staff Progress Report*. US Department of the Treasury, Board of Governors of the Federal Reserve System, Federal Reserve Bank of New York, US Securities and Exchange Commission, and US Commodity Futures Trading Commission.

Kutai, Ari, Daniel Nathan, and Milena Wittwer. 2025. "Exchanges for Government Bonds? Evidence during COVID-19." *Management Science*. https://doi.org/10.1287/mnsc.2023.02344.

Krishnamurthy, Arvind. 2010. "How Debt Markets Have Malfunctioned in the Crisis." *Journal of Economic Perspectives* 24 (1): 3–28.

Li, Dan, Lubomir Petrasek, and Mary Tian. 2024. "Risk-Averse Dealers in a Risk-Free Market—The Role of Internal Risk Limits." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4824250.

Liang, Nellie. 2022. "Remarks by Under Secretary for Domestic Finance Nellie Liang at the 2022 Treasury Market Conference." Speech, US Department of the Treasury, November 16. https://home.treasury.gov/news/press-releases/jy1110.

Metrick, Andrew. 2024. "The Failure of Silicon Valley Bank and the Panic of 2023." *Journal of Economic Perspectives* 38 (1): 133–52.

Rigon, Lorenzo. 2024. "Three Essays on Central Banking and Debt Management." PhD diss., Stanford University.

SEC. 2025. "SEC Extends Compliance Dates and Provides Temporary Exemption for Rule Related to Clearing of U.S. Treasury Securities." Press release, February 25. https://www.sec.gov/newsroom/press-releases/2025-43.

Tarullo, Daniel K. 2023. *Capital Regulation and the Treasury Market*. Hutchins Center on Fiscal and Monetary Policy, Brooking Institution.

Tobin, James. 1996. "An Essay on Principles of Debt Management." In *Handbook of Debt Management*, edited by Gerald J. Miller, 693–736. Routledge.

Vayanos, Dimitri, and Jean-Luc Vila. 2021. "A Preferred-Habitat Model of the Term Structure of Interest Rates." *Econometrica* 89 (1): 77–112.

Weiss, Colin. 2022. "Geopolitics and the US Dollar's Future as a Reserve Currency." Board of Governors of the Federal Reserve System International Finance Discussion Paper 1359.

Weiss, Colin R. 2022. "Foreign Demand for US Treasury Securities during the Pandemic." FEDS Notes, January 28. https://www.federalreserve.gov/econres/notes/feds-notes/foreign-demand-for-us-treasury-securities-during-the-pandemic-20220128.html.

Wessel, David. 2024a. "How to Tell if the US Treasury is Having Trouble Borrowing in the Bond Market." Brookings Institution, July 23. https://www.brookings.edu/articles/how-to-tell-if-the-us-treasury-is-having-trouble-borrowing-in-the-bond-market/.

Wessel, David. 2024b. "What Is Bank Capital? What Is the Basel III Endgame?" Brookings Institution, March 7. https://www.brookings.edu/articles/what-is-bank-capital-what-is-the-basel-iii-endgame/.

Williams, John. 2021. "Preparing for the Unknown." Speech, 2021 US Treasury Market Conference, November 17. https://www.newyorkfed.org/newsevents/speeches/2021/wil211117.

Yadav, Yesha and Josh Younger. 2025. "Central Clearing in the US Treasury Market." Vanderbilt Law Research Paper 25-01.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 215–234

US Corporate Bond Markets: Bigger and (Maybe) Better?

Maureen O'Hara and Xing (Alex) Zhou

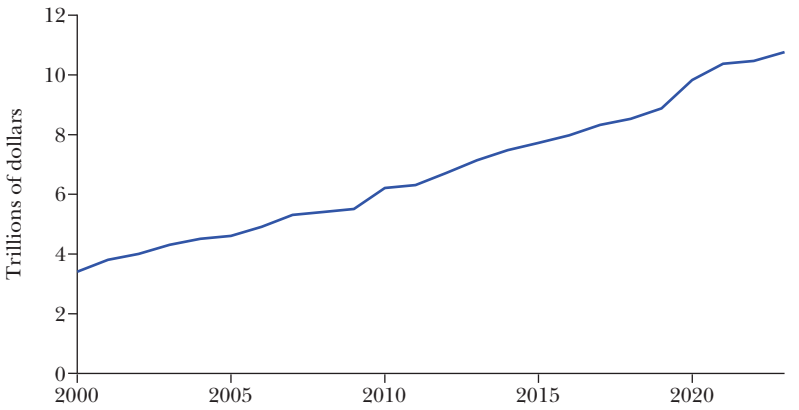
Corporate bonds are "fixed-income securities" that provide investors a promised fixed yearly payment during the life of the bond and the repayment of the principal (or investment amount) at maturity. Unlike owning a stock, whose return is uncertain but has the potential for a large upside return, a bond promises a certain annual payment but gives up any upside appreciation. For US corporations, the certainty of payments and longer maturities have long made corporate bonds the lifeblood of corporate borrowing. Companies borrow to finance a variety of activities such as capital expenditures, research and development, mergers and acquisitions, and new initiatives. Bonds provide the longer-term fixed-rate financing to facilitate these endeavors in contrast to bank loans which are normally of shorter maturities and often feature variable interest rates. As Figure 1 shows, the US corporate bond market is a behemoth: with over $11 trillion in outstanding bond issues, it far exceeds the $2.8 trillion of outstanding commercial and industrial loans at commercial banks. For a discussion of trends in corporate bond and loan markets, Bochner, Wei, and Yang (2020) offer a useful starting point.

The corporate bond market is changing rapidly. Primary issuance has exploded, creating an ever-larger market. The buyers in bond markets have long been dominated by institutional investors like insurance companies and pension funds, but retail investors are now entering the market in greater volume, attracted by (or

■ *Maureen O'Hara is Robert W. Purcell Professor of Management, Johnson College of Business, Cornell University, Ithaca, New York. Xing (Alex) Zhou is Associate Professor of Finance, Cox School of Business, Southern Methodist University, University Park, Texas, and a Research Fellow, Office of Financial Research, US Department of the Treasury, Washington, DC. Their email addresses are mo19@cornell.edu and axzhou@smu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20251439.

Figure 1
Total Outstanding Amount of US Corporate Bonds



Source: SIFMA.
Note: This figure shows obligations of US financial and nonfinancial corporations including bonds, notes, debentures, mandatory convertible securities, long-term debt, private mortgage-backed securities, and unsecured debt. Includes bonds issued both in the United States and in foreign countries, but not bonds issued in foreign countries by foreign subsidiaries of US corporations. Recorded at book value.

perhaps causing) the rapid growth of bond mutual funds and exchange-traded funds. Corporate bond trading is shifting to electronic markets, adding competition to the traditional dealer-based model of liquidity provision. New entrants such as proprietary trading firms are playing roles akin to the high frequency traders found in other markets.

Whether this "new" corporate bond market is actually "better" depends on the answer to a simple question: Can this market structure handle the age-old problem of illiquidity in corporate bond trading? Illiquidity means that an asset cannot be easily sold without suffering a large fall in price. Typically, bonds were purchased by institutional investors, and then often held until maturity, or resold only very infrequently. Thus, while buying a bond was easy, subsequently selling the bond before maturity was not. In times of financial stress, if a number of bond-holders wanted or needed to sell bonds at the same time, the secondary market for buying these bonds was quite limited. It was possible for a "fire sale" dynamic to emerge in which illiquidity in the corporate bond market meant that the bonds could only be sold at a substantial loss. With this scenario in mind, certain features of the new bond market are troubling: market growth may be outstripping liquidity provision, electronic liquidity provision is more erratic, and market linkages may introduce episodic instability. Moreover, the Federal Reserve provided an unprecedented bailout of the US corporate bond market during the COVID-19 pandemic, which may have led to moral hazard in a way that diminished attention paid to these concerns.

Our premise in this paper is that the new corporate bond market is better in many ways, but it is more fragile. We argue that an expensive but stable market

structure for corporate bonds has given way to a market that is more efficient, but less resilient. Focusing on three major changes in the bond market—dramatically lower transaction costs, reduced search costs, and greater price transparency—we delineate how this evolution is providing a variety of benefits to investors and firms. However, we show that these benefits accrue only in normal market times; they can be noticeably absent in the stress periods when liquidity is most needed. We argue that this episodic illiquidity reflects a shift from corporate bond trading being a relationship-based business to a more transaction-based business. Given the central role the corporate bond market plays in the overall economy, this new instability raises additional concerns about systemic risk in the financial system.

One aspect of this bond market evolution that we find particularly intriguing is how it highlights the complex and sometimes conflicting roles that markets play. Greater competition can lower trading costs, but it can also undermine the willingness of market intermediaries to provide liquidity by buying the bonds in challenging times. As we discuss, most corporate bonds are inherently illiquid, but few of the changes in the market address the fundamental problem of who is going to provide liquidity when it is particularly needed. We suggest some alternatives for how to enhance liquidity provision in this ever-bigger corporate bond market.

A Brief Introduction to Corporate Bond Markets

Corporate bond issues come to the market via an underwriting process; that is, investment banks agree, for a fee, to arrange the sale of the bonds to institutional investors. Bonds are generally issued in very large denominations, so retail investors play a small role in this initial primary market. There are two main categories of corporate bonds: investment grade and high yield. What determines which category a particular bond falls into is its probability of default and the potential severity of losses in the event of default, as predicted by a credit rating agency. For example, S&P (formerly Standard & Poor's) ranks bonds from AAA (the best credit) to D (the bond is already in default). Investment grade bonds have ratings of BBB− or above; high yield bonds have ratings below that level.

Corporations can have multiple issues of outstanding bonds, delineated by their final maturity date. For example, Ford Motor Company currently has 30 outstanding bond issues. In the US market overall, there were 89,913 outstanding corporate bond issues coming from 4,474 issuers in summer 2024. Because of the plethora of bond issues, many bonds rarely trade, generating the illiquidity issue noted earlier. Investors in corporate bonds were traditionally institutional investors such as insurance companies, pension funds, endowment funds, and the like (for a discussion of ownership of corporate bonds, see Koijen and Moto 2023). These long-term investors purchased bonds in the primary market and often held the bonds until maturity. Indeed, insurance companies are still the largest holders of corporate bonds, holding approximately 30 percent of all outstanding corporate bonds. More recently, corporate bonds have found demand coming from mutual

funds and exchange-traded funds. These investment vehicles face customer inflows and outflows, necessitating purchases and sales of bonds in the secondary market—that is, the market for bonds that have already been issued.

Secondary-market bond trading (or trading of bonds that have already been issued) has been handled in a dealer market. Unlike the trading of equities, which takes place on central exchanges like the New York Stock Exchange, dealer markets are decentralized, with customers having to contract individual dealers (typically by phone) to get their quotes to buy or sell a particular bond issue. These bond dealers include the banks who brought the issues to the primary market, but many other dealers make markets in bonds as well. O'Hara and Zhou (2022) report over 1,000 corporate bond dealers in 2012, although that number is declining rapidly (for reasons we discuss in the next section), having fallen to just over 500 dealers in 2022.

With hundreds of dealers, customers face search costs in finding the best prices for a particular corporate bond. For actively traded issues, large dealers—typically investment banks—may hold inventories of such bonds (or, given their primary market activities, have better information on where to find the bonds). Other issues may be smaller or less-frequently traded, making it harder to locate counterparties. These search costs, as well as market power on the part of the larger dealers, have resulted in large transaction costs in secondary bond trading. An extensive academic literature examines the behavior of dealer-market bond trading (for example, see Duffie, Gârleneau, and Pedersen 2005; O'Hara, Wang and Zhou 2018; Goldstein and Hotchkiss 2020).

A related literature examines the strategies of customers in this dealer-based world. Certain institutional customers also can have market power in corporate bonds for two reasons: dealers make profits from repeat customer trades and dealers also learn useful market pricing information from customer orders. To exploit these features of bond trading, relationship networks develop in which customers concentrate trading with certain dealers and dealers give such customers "better" executions. Using insurance company data, O'Hara, Wang, and Zhou (2018) show how insurers trading with a smaller network of dealers improve their execution quality (that is, they are able to buy at lower prices and sell at higher prices). Indeed, one-third of insurance companies concentrate their bond trades with a single dealer (Hendershott et al. 2020).

One other intriguing market pattern we highlight here is the changing composition of issuers. Over time, the number of issuers has remained relatively stable, with 4,694 corporate bond issuers in 2001 and 4,474 issuers in 2023. The largest issuers in 2001 included a variety of nonfinancial firms, with some of the most active including Kroger, Ford, CMS Energy, and General Motors. Looking at all issuers in 2001, the average number of outstanding bonds per issuer was five. By 2023, the market looked very different. Now the 20 largest issuers are all financial firms, with JPMorgan Chase Financial, Citigroup, Goldman Sachs, UBS, Wells Fargo, and Morgan Stanley actively raising funds through multiple bond issues. Indeed, JPMorgan Chase Financial alone now has 10,000 outstanding bond

issues, a staggering number compared to 116 outstanding bond issues of Chase Manhattan Corp. in 2001. What is driving this compositional shift is unclear, but new leverage requirements on banks as well as stress tests for banks required by the Wall Street Reform and Consumer Protection Act of 2010—commonly known as the "Dodd-Frank Act"—may at least partially explain the increased bond issuance by banks. Whatever the cause, this increasing dependence of the financial firms on the bond market highlights the growing interdependency of the overall financial markets. It also makes clear how economic stress in the bond market has the potential to translate into economic stress in the banking sector, something that was less likely when bonds were primarily issued by nonfinancial corporations. We discuss the implications of this change further in the conclusion.

The New World of Bond Trading

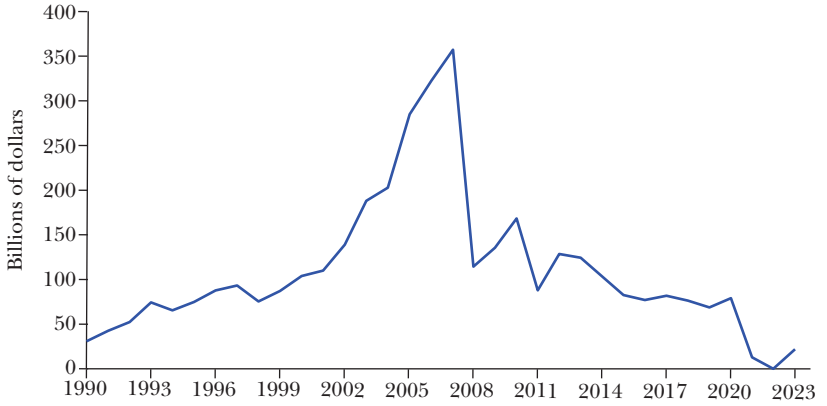
The Diminishing Role of Dealer Trading

The corporate bond market has relied on dealers, like those in investment banks, for providing liquidity. Given the over-the-counter nature of the bond market, it is challenging for investors to search for a counterparty and to arrange trade directly with each other. Dealers provide liquidity to the market by committing their capital to buy what investors wish to sell and to sell what investors wish to buy. Unlike in the equity market, trades in the bond market tend to be for large sizes and arrive at irregular times. As a result, the capability of dealers to warehouse investor flows on their balance sheets is essential to the proper functioning of the market.

Prior to the 2008 global financial crisis, many dealers, particularly large banks, were ready to expand their balance sheets to accommodate customer bond trades. Along with the growth of the corporate bond market, dealers increased their bond inventories rapidly from around $100 billion in 2000 to over $350 billion in 2007, as shown in Figure 2. But the financial crisis revealed solvency and liquidity issues for many banks. Regulators adopted a number of reforms, including the US Dodd-Frank Act and Basel III international banking regulations, to raise bank capital requirements and increase liquidity. While these regulatory reforms arguably made banks safer and increased the resilience of the financial system, they also increased the cost of capital and restricted risk-taking in a way that made banks less willing to hold large inventories of corporate bonds. In particular, the "Volcker rule," enacted as part of the Dodd-Frank Act, prohibits banking entities from engaging in proprietary trading.

As banks are the primary dealers for many fixed income securities like bonds, the rule exempts their market-making activities. However, distinguishing legitimate market-making from prohibited proprietary trading can be challenging (Duffie 2012). In illiquid markets such as the corporate bond market, the rule disincentivizes dealers from taking large positions that cannot be easily unwound, as slow turnover of inventory may raise concerns for regulators that the dealer is engaged

Figure 2
Dealer Corporate Bond Position



Source: Federal Reserve Board, "Financial Accounts of the United States."
Note: Corporate bond positions include domestic and foreign bonds held in the United States by securities broker-dealers.

in proprietary trading. Dealer reluctance to provide liquidity can be detrimental to the bond market, especially in stress times when demand for liquidity heightens. Indeed, when Bao, O'Hara, and Zhou (2018) study a sample of stressed bonds that were downgraded from investment grade to speculative grade, they find that after the Volcker rule, liquidity in these markets deteriorated, and dealers committed less capital and prearranged more trades when intermediating the trading of these bonds.

In addition to the Volcker rule, higher capital requirements for banks and other financial firms discourage provision of liquidity to the bond market by holding corporate bonds. Adrian, Boyarchenko, and Shachar (2017) find that dealers facing more regulations after the financial crisis of 2008–2009 intermediated fewer customer trades. They construct measures of dealer-level balance sheet constraints, and show that liquidity declined more in bonds intermediated by dealers with more constrained balance sheets. The potential impact of post-crisis regulations on corporate bond market liquidity has been studied in a large number of papers, with most documenting a deleterious effect (for example, Bessembinder et al. 2018; Choi, Huh, and Shin 2024; Dick-Nielsen and Rossi 2019; Schultz 2017). Consistently, Figure 2 shows that dealers' aggregate positions in corporate bonds have continued to decline after the financial crisis, and approached zero in 2022 during the recent monetary policy tightening cycle.

In addition to regulatory constraints, dealer funding costs have also played an important role in driving bond liquidity in the years since the global financial crisis of 2008–2009. Despite improved levels of capital, costs for bank dealers to fund their bond inventories are substantially higher than their pre-crisis levels. For example,

Berndt, Duffie, and Zhu (2025) estimate roughly 170 percent higher wholesale debt financing costs for large banks after controlling for insolvency risk. Higher funding costs, in turn, discourage dealers from growing large balance sheets (Andersen, Duffie and Song 2019). Macchiavelli and Zhou (2022) show that dealers' funding costs in repurchase markets affect their liquidity provision in the corporate bond markets.

The Rise of Electronic Trading

Against the backdrop of a rapidly growing corporate bond market, but with reduced liquidity provision by dealers, corporate bond trading has started to migrate from traditional voice-based venues to electronic platforms, such as MarketAxess, TradeWeb, and Bloomberg. A variety of trading protocols have developed on these platforms to facilitate the matching of buyers and sellers. To date, the most popular protocol is called request-for-quote (RFQ), a protocol which essentially kept a role for dealers but reduced the costs of finding counterparties for trades. On an RFQ platform, a customer sends an inquiry for trading a single bond or a list of bonds to a number of dealers with whom the customer has an existing permissioned trading relationship. The inquiry typically includes customer identity, bond identity, the side (buy or sell), size (the amount intended to trade), respond time window (for example, ten minutes for investment-grade bonds) and other relevant parameters. Based on the responses received, the customer can select a dealer to trade and the trade is completed. Importantly, in an RFQ, dealers know neither the number or identities of the other dealers contacted, nor do they know the quotes other dealers provide to the customer.

The growth of request-for-quote trading has brought greater liquidity and efficiency to the corporate bond market. Hendershott and Madhavan (2015) combine data from TRACE, the Trade Reporting and Compliance Engine developed by government regulators to require that dealers report every over-the-counter corporate bond transaction, with data from MarketAxess, a leading RFQ trading platform, and analyze a sample of corporate bond transactions in US corporate bonds from January 2010 through April 2011. They argue that RFQ increases dealer competition, reduces customers' search costs, and results in better trading prices for customers. Even after accounting for customers' endogenous choices about trading venue, transaction costs tend to be lower on an RFQ platform. Similarly, in a study of the evolution of bond electronic trading over the period from 2010 to 2017, O'Hara and Zhou (2021a) show that electronic trading has had a wide-ranging impact on the corporate bond market. The proliferation of electronic trading platforms has changed dealer behavior. In addition to increasing competition among dealers, electronic trades facilitate dealers who are searching for counterparties when managing their own inventory risk. One immediate impact of this change has been a dramatic decline in the volume transacted in the inter-dealer market, the traditional method dealers used to offset unwanted exposures. Now dealers, like customers, can use these electronic platforms to adjust undesired inventory positions. Electronic trading also provides more information that allows dealers to set better prices. As a result, the growth of electronic trading has also led to a decline

in transaction costs for both trades executed electronically and for traditional voice trades.

However, the benefits brought by request-for-quote trading have not been universal. The decrease in transaction costs is mostly observed in more liquid issues, such as larger investment-grade bonds and those bonds more recently issued or with a short time to maturity (Hendershott and Madhavan 2015). Beneficial effects were also limited to smaller-sized trades or those with lower cost of information leakage. Similarly, O'Hara and Zhou (2021a) also find that small trades seem to benefit more from electronic trades; that is, the pattern of higher transaction costs for smaller sized trades, typically observed in corporate bond trading, disappeared in RFQ trades. By the end of their sample period, transaction costs for all RFQ trades appear to be converging to 10 basis points for investment-grade trade bonds and to 20 basis points for high-yield bonds. In addition, the cross-venue effects of electronic trades on traditional voice trades have also been limited to smaller-sized categories.

The limitations of request-for-quote trade can at least partially be attributed to the way the market is set up. The protocol only allows customers to approach dealers with whom the customer already has an existing trading relationship—and hence still reflects the bilateral nature of the previous over-the-counter trading. Therefore, even though RFQ improves trade execution along several dimensions, dealers remain the main liquidity providers on RFQ platforms. O'Hara and Zhou (2021a) use data from MarketAxess to identify RFQ trades in the TRACE data on over-the-counter transactions collected by the federal government. Using the dealer identities behind each trade included in the regulatory version of the TRACE data, they show that RFQ electronic trading is dominated by almost the same dealers that intermediate most traditional voice trading. Nine out of the ten largest dealers rank in the top 15 dealers in both voice-trading and RFQ electronic trading. Kozora et al. (2020) also show that the vast majority of trading on alternative trading system platforms executing trades using RFQ protocols involve dealers. Given the increased reluctance of dealers to provide liquidity during the era after the global financial crisis of 2008–2009, this continued heavy reliance on dealers for liquidity provision in RFQ trading raises an important question: While electronic trading has made trading more efficient, does it actually address the liquidity changes in the corporate bond market?

To encourage broader participation in liquidity provision, especially by "buy-side" institutions like mutual fund complexes or private equity firms that want to purchase assets as investments, an alternative trading protocol called all-to-all trading has been introduced to the corporate bond market. All-to-all trading essentially enables any market participant to trade directly with any other market participant. Conceptually, such trading protocols can help improve bond market liquidity when dealers' intermediation capacity is constrained, because a buy-side institution can execute a trade directly without another buy-side institution without involving a dealer. In 2012, MarketAxess started Open Trading to incorporate all-to-all trading into its existing request-for-quote trading platform. Open Trading increases competition in RFQs and helps improve prices for investors (Hendershott,

Livdan, and Schürhoff 2021). However, the growth of all-to-all trading has been fairly slow. By 2018, Open Trading accounted for 12 percent of trade volume on MarketAxess, which in turn represents about 10 percent of the total corporate bond market trade. More importantly, the majority of Open Trading still involves a dealer. Of the 12 percent of Open Trading transactions, only 2 percent were executed directly between investors without dealers.

The slow growth of investor-to-investor trading might reflect the lack of pricing capabilities by many buy-side institutions in the corporate bond market. Many bonds feature covenants (or restrictions) that make their value in various future states of the world quite complicated. For example, some bonds are callable, meaning that the issuer has the option to buy the bond back under certain conditions. Pricing that embedded option, and thus the value of the bond, need not be straightforward. Similarly, if an issuer enters bankruptcy, the entire balance sheet of the corporation becomes relevant, so that pricing a bond with even a small chance of default can be very complex. While intermediaries will have this requisite pricing knowledge, it may be beyond that of other investors for all but the largest, most actively traded issues.

Overall, electronic trading has helped improve liquidity in the corporate bond market. However, the robustness of liquidity provided through electronic trading platforms is a concern. Focusing on the time periods around rating downgrades where there is increased demand for liquidity, O'Hara and Zhou (2021a) show that trading shifts from request-for-quote platforms back to voice trading, consistent with traders relying more on dealer relationships rather than on electronic trading to source liquidity. In addition, the advantage of lower transaction costs for RFQ trading, observed in normal times, disappeared following bond rating downgrades. Electronic trading costs rise higher than voicing trading costs.

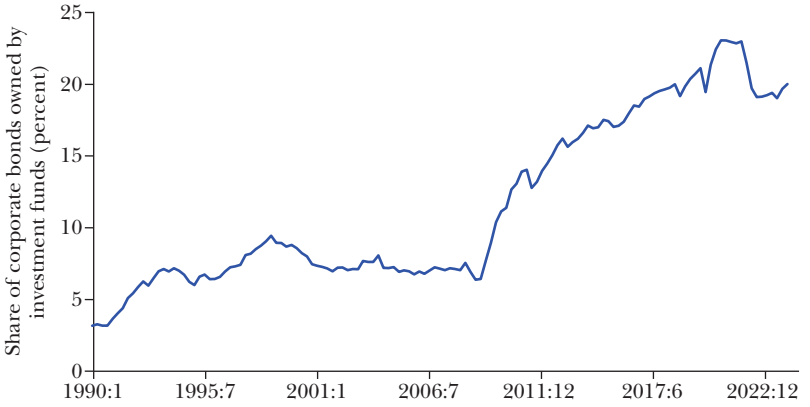
The fragility of liquidity provided through electronic platforms is also observed in all-to-all trading. During the COVID-19 liquidity crisis in the corporate bond market, electronic all-to-all trade volume did increase, but it remained a small share of the overall bond market (O'Hara and Zhou 2021b). More importantly, transaction costs for investor-to-investor trading, which were lower than those involving a dealer before the crisis, more than doubled those between investors and dealers during the crisis.

One explanation for this faltering of electronic trading during financial crises may lie in the nature of relationship versus transactional trading. Electronic trading is more transactional; buyers and sellers are matched for this transaction only. In contrast, dealer markets are more relationship-based, with the promise of future business a factor in a dealer's pricing decisions. That trading shifts from electronic markets to dealers in times of financial stress suggests that relationships may be a source of liquidity when electronic counterparties are scarce.

Market Linkages from Corporate Bond Mutual Funds

Retail investors have been discovering corporate bonds as an investment class, attracted perhaps by their higher yields relative to other investments like bank

Figure 3
Investment Fund Ownership of Corporate Bonds



Source: Federal Reserve Board, "Financial Accounts of the United States."
Note: The figure plots the fraction of corporate and foreign bonds (held in the United States) owned by mutual funds and exchange-traded funds.

certificates of deposits. Unlike institutional investors, which generally hold individual bonds, retail investors access this asset class via mutual funds and exchange-traded funds. While institutions still hold the lion's share of corporate bonds, the growth of investment fund holdings has been impressive. As Figure 3 shows, in January 2009 investment funds held only 6 percent of outstanding bonds; by January 2024 their holdings stood at 20 percent, down slightly from their all-time high of 23 percent in October 2021.

Increased demands for corporate bonds from investment funds should have positive effects for bond issuers and investors alike. Retail investors get increased access to an asset class whose market characteristics (large denomination, dealer market trading, and so on) hitherto discouraged investment. Issuers benefit from increased demand, raising bond prices and so lowering yields. But the actual impact on the corporate bond market is more nuanced. Bond funds essentially turn an underlying illiquid asset into a liquid investment. Demands for liquidity if there is selling at the fund level can then turn into demands for liquidity at the bond level. This has the potential to increase bond market fragility.

To see why, consider the effect of investor inflows and outflows in mutual funds. Funds allow investors to invest and redeem on a daily basis. It has long been established that equity mutual fund flows have a convex relationship with fund performance—that is, when the mutual fund is performing better, fund inflows increase; when the fund performs worse, fund outflows also increase, but not by as much. However, this convexity does not seem to be a feature of corporate bond mutual funds. Goldstein, Jiang, and Ng (2017) found that under some performance metrics, outflows from corporate bond funds were more sensitive to bad

performance than inflows were to good performance. Large demands to sell the liquid mutual fund can result in large price movements in the underlying illiquid asset (the bond). So, if more investors in a corporate bond mutual fund want out in bad times, it behooves investor to get out sooner than later—or what is known as the first-mover advantage. This behavior sets the stage for destabilizing "fire sales," in which prices can be driven down quickly in the short-term in both mutual funds for corporate bond and the underlying bonds themselves.

A large literature has investigated the importance of this effect on the underlying corporate bond market. On one side, using data from 2002 to 2014, Choi et al. (2020) argue that this effect is minimal. They argue that funds handle redemption risk by holding cash reserves and often Treasury securities, and these assets are the first sold to meet redemptions rather than the underlying bonds. When corporate bonds are sold, they find that these are the most liquid bonds, which also should have lessened any impact on the bond market.

More recent research, however, reaches different conclusions. The COVID-19-related financial crisis caused widespread stress in bond markets. Treasury, corporate bond, and municipal bond markets all suffered difficulties as trading volumes soared, volatility spiked, and liquidity evaporated. Bond mutual fund redemptions in March 2020 reached 5 percent, a level never before seen. Using data from this time, Ma, Xiao, and Zeng (2022) provide strong evidence that these mutual fund outflows negatively affected underlying bond prices. These authors, like Choi et al. (2022), found that mutual funds followed a pecking order for selling assets to meet redemptions, first unloading Treasury securities, then higher rated corporate bonds, and finally lower-rated bonds. But unlike Choi et al. (2020), the study by Ma, Xiao, and Zeng (2022) finds that these fund management techniques were not enough to offset the effects of fund flows on the underlying market. These authors estimate that in the Covid crisis investment-grade corporate bond yields would have been 42 basis point lower in the absence of mutual fund ownership. Related research by Haddad, Moreira, and Muir (2021) reaches a similar conclusion that selling pressure from mutual fund flows in the COVID-19 crisis was a major factor explaining the dramatic meltdown in the corporate bond market.[1] Overall, research supports the conclusion that market linkages introduced by bond mutual funds are now an important factor in making fixed income markets less resilient in the face of stress conditions.

Market Linkages from Corporate Bond Exchange-Traded Funds

Corporate bond exchange-traded funds also engage in transforming underlying illiquid assets into liquid investment vehicles. These funds hold relatively

[1] Li, O'Hara, and Zhou (2024) reached similar conclusions based on evidence from the municipal bond markets. The growth of muni-bond mutual funds is an even more recent phenomena than corporate bond mutual funds and exchange-traded funds. Municipal bonds are also more illiquid than corporate bonds and these bond funds hold less in the way of liquid reserves. These authors found that dealers reduced their inventories, liquidity worsened, and yield spreads reflected a fire sale premium for bonds held by mutual funds.

illiquid corporate bonds, but trade on highly liquid exchanges like the New York Stock Exchange or the NASDAQ. Unlike bond mutual funds, bond investors in exchange-traded funds sell their fund shares directly on these exchanges and so do not feature daily redemptions. However, exchange-traded funds do have a unique arbitrage mechanism through which shocks to the fund can spill over to the underlying corporate bond market. When the market price of an exchange-traded fund rises above or falls below its net asset value, the "authorized participants" for the exchange-traded fund, typically large maker makers, exchange a representative portfolio of corporate bonds (a "creation basket"/ "redemption basket") for shares in the exchange-traded fund. Such trades will drive down or push up prices of the exchange-traded fund, and close the arbitrage gap.

Over the past two decades, corporate bond exchange-traded funds have experienced significant growth. The total assets under management by US corporate bond exchange-traded funds has risen from roughly $4 billion in 2007 to $300 billion by August 2024.[2] The growing significance of corporate bond exchange-traded funds have raised concerns about their potential impact on liquidity of the corporate bond market, especially when the market is under stress.

In a study of a sample of corporate bond exchange-traded funds and their security-level holdings over the period from January 2009 to November 2013, Dannhauser (2017) finds that the ownership of exchange-traded funds has no significant impact on high-yield corporate bonds, but has a negative impact on investment-grade bonds. She argues that the negative liquidity effect can be attributed to the liquidity mismatch between exchange-traded funds and the underlying corporate bonds, which encourages liquidity trades to migrate from the illiquid bond market to highly liquid exchange-traded funds, and thus reduces the liquidity of the underlying bonds. However, Holden and Nam (2024) argue that impact of bond exchange-traded funds on the underlying bonds depends on the accessibility of the bond market. When bond market accessibility is low, the introduction of a bond exchange-traded fund can attract new uninformed traders, and the liquidity of the exchange-traded fund can spill over to the underlying bonds by arbitrage trading. They compare liquidity of underlying corporate bonds before and after the introduction of exchange-traded fund, and find that liquidity improvements are larger for bonds less accessible to investors (that is, highly arbitraged, low volume, high-yield, long-duration bonds, and so-called "144A" bonds that are issued in a private placement to Qualified Institutional Buyers).

Some research directly links bond liquidity to the creation and redemption processes of exchange-traded funds and examines the linkage both in normal and stress times. Finnerty, Reisel, and Zhong (2024) use data on both reported and realized baskets and find that including a bond in a creation or redemption basket improves bond liquidity both for high-yield and investment-grade markets. They show that that the benefit of bond exchange-traded funds exists in both normal and stress times, and

[2] See the VettaFi database for corporate bond exchange-traded funds at https://etfdb.com/etfdb-category/corporate-bonds/.

argue that such benefit can be explained by mispricing arbitrage. In contrast, Koont et al. (2023) argue that the impact of active portfolio management on bond liquidity differs depending on market conditions. Corporate bond exchange-traded funds actively manage their bond portfolios by balancing index-tracking against liquidity transformation. In normal times, a particular bond can be randomly included in creation or redemption baskets across exchange-traded funds. This, in turn, increases the bond's trading activities and improves its liquidity. However, in stress times, large redemptions move bonds from exchange-traded funds to the balance sheets of authorized participants for that fund. Because most authorized participants are also bond dealers and they use their balance sheets in both roles (Pan and Zeng 2019), they tend to reduce liquidity provision in the same bonds. Koont et al. (2023) analyze the COVID-19 crisis period and provide empirical findings that support this argument.

One of the key challenges in studying the impact of exchange-traded funds on the underlying bond market is that the inclusion of a bond in exchange-traded fund baskets, especially in stress times, is endogenous. Indeed, some market participants argue for reverse causality: "it is because some bonds are illiquid that they increasingly feature in redemption baskets as sell-offs intensify, not vice versa" (as reported by Johnson 2023). Consistent with this view, Todorov (2021) argues that when facing runs, sponsors of exchange-traded funds adjust redemption baskets to include riskier or less-liquid bonds. In sum, exactly how exchange-traded funds affect the liquidity and stability of the corporate bond market requires further study, but what is clear is that bond exchange-traded funds do influence the underlying bond market, for better or worse.

Making the Bond Market More Resilient

As the corporate bond market grows ever larger, the problems highlighted in this discussion will take on even greater importance. More bonds (and more trading) require more liquidity, but corporate bonds are inherently illiquid. Thus, the shift towards a more transaction-based market structure leaves unanswered the basic question of who will buy when everyone else is selling? In this section, we explore three mechanisms to address the underlying challenge of episodic illiquidity in bond markets.

The Federal Reserve Backstop

The central bank has long played a "lender of last resort" role of backstopping the banking system in times of economic distress. Banks basically lend for long maturity and borrow for shorter maturities, meaning that at any point in time, they are essentially illiquid. Acting as a lender of last resort, the central bank can step in to supply liquidity directly when banks are facing a liquidity crisis due to increased demand to withdraw funds. This central bank support has not generally been extended to the fixed income sector of the financial markets. However, this changed during the recent COVID-l9 financial crisis, in March 2020, when the Fed

took the momentous step of bailing out the corporate bond market. (Our discussion here draws heavily on O'Hara and Zhou 2021b).

To understand what happened and why, we start with Silbert and Buiter's (2007) observation that in modern markets, where bonds and other fixed income markets like US Treasury securities are larger than banking markets, a liquidity crisis arises in a different way than in the lender-based problems of the past. Instead, a liquidity crisis results in disorderly markets in which "there is no market maker with both the knowledge and deep pockets to credibly post buying and selling prices." To address such a liquidity crisis, the authors argue the central bank must act as a "market maker of last resort," being willing to buy assets themselves or to provide the financing to allow others to do so. This role, very much akin to the lender of last resort function, positions the central bank as a backstop when all other sources of liquidity are scarce or nonexistent.

Starting in late February 2020, the COVID-19 crisis caused financial markets worldwide to face unprecedented demands for liquidity. In US financial markets, this manifested in extreme selling pressure across fixed-income markets: Treasury bond, corporate bond, and municipal bond markets. This desire to sell caused corporate bond prices to deteriorate rapidly after March 6, with transaction costs for top-rated bonds increasing to 90 basis points, triple their level from the month before. Trading for large size trades were particularly hard hit, with transaction costs exploding from 24 basis points to more than 150 basis points by March 23. Adding to this tumult were huge outflows from bond mutual funds and exchange-traded funds, fueling further demands for liquidity in the underlying bond market.

Everywhere there were eager sellers; buyers, not so much. As the crisis progressed, bond dealers shifted from buying bonds to selling bonds, resulting in a negative $8 billion inventory position for the dealer community. Thus, the dealers, rather than supplying liquidity, turned to demanding liquidity, further exacerbating the liquidity crisis. Indeed, bonds sold more aggressively by dealers experienced a greater increase in transaction costs than other traded bond issues.

Other potential sources of buy-side liquidity for bonds did not step forward. Electronic trading platforms held out the promise of both dealer-to-customer trading and customer-to-customer trading. Volume in customer-to-customer electronic trading did increase, but it remained small and extremely expensive at more than double the cost of customer-to-dealer trading. Moreover, it became impossible to execute more than the smallest trade sizes, providing little in the way of liquidity to the army of those wishing to sell. With neither dealers nor customers seemingly willing to provide liquidity, the bond market appeared to be heading for free fall.

The Federal Reserve responded with a variety of actions. On March 17, 2020, the Fed announced the creation of the Primary Dealer Credit Facility (PDCF), which offered direct loans to "primary dealers." Primary dealers are a subset of dealers, generally the large dealers at major banks, who play a fundamental role in both the Treasury bond market and the corporate bond market. The PDCF offered overnight and term funding up to 90 days for loans collateralized by investment-grade bonds. Lending money to dealers to buy bonds is akin to the lender

of last resort role traditionally played by the Fed, albeit now in a broader market setting. More groundbreaking was the action announced on Monday, March 23, when the Fed, together with the Department of the Treasury, created the Secondary Market Corporate Credit Facility (SMCCF) to directly purchase investment-grade corporate bonds in the secondary market (this was later extended to purchasing exchange-traded funds). Here, the Fed explicitly took on the "market maker of last resort" role envisioned by Buiter and Silber (2007), adding corporate bonds to the Federal Reserve's balance sheet.

Why was this second program necessary? While the Primary Dealer Credit Facility provided funding to dealers, the threat of continued excessive pressures for selling in the bond market meant that dealers lacked an incentive to add to inventories—irrespective of funding concerns. The Secondary Market Corporate Credit Facility offered a backstop to dealers, providing a willing buyer to help offload inventories if the need arises.

The Fed's actions to stabilize the corporate bond market were generally successful. Particularly after the launch of the Secondary Market Corporate Credit Facility, and as the uncertainties surrounding the COVID-19 epidemic began to abate, bond yield spreads declined, although they would remain substantially higher than their pre-crisis levels for many months. By providing liquidity backstops for bond holdings of mutual funds, SMCCF reduced investors' incentives to run and reversed fund outflows, particularly for the most fragile bond funds (Falato, Goldstein, and Hortaşçu 2021).

By acting as market maker of last resort, the Federal Reserve provided one answer in March 2020 to the question, "Who will buy bonds when everyone is selling?" But taking this step raises a variety of other equally important questions about the functioning of the bond market. For example, if the Fed rides to the rescue in the future, what incentive is there for dealers or others to provide liquidity in difficult markets? This moral hazard problem raises the specter that periodic instability may potentially become more frequent, rather than less so. A related concern is that the Fed was only willing to buy specific types of bonds (initially, for example, only corporate investment grade) and maturities (facilities were focused on maturities of five years or less). Could these Fed restrictions affect the assessing and pricing of credit risk, and ultimately the allocation of credit in the economy? Time will tell, but it seems clear that the current structure of the bond market is less resilient than the traditional relationship-based market of times past. The Fed may need to play a continuing role to ensure the stability of this evolving market structure.

The Role of Long-Term Investors

Having the Federal Reserve step in as market maker of last resort may be the only realistic solution in a market meltdown. But can the corporate bond market be made more resilient so that such a point is less likely to arise? As we have discussed, a variety of factors limit the role that dealers now play in the corporate bond markets. Moreover, dealers have a short horizon with respect to liquidity provision; they

are willing to buy and sell because they expect to be able to reverse their position at a profit in the near future. What adds resiliency to a market are investors with long-term horizons who can step in and buy when illiquidity and disruption in the market offer opportunities to invest at bargain prices.

The corporate bond market has a variety of such long-term investors. Endowments and pension funds, for example, have long horizons. Particularly important are insurance companies who strive to match their long-term liabilities with long-term assets such as corporate bonds. As the largest holders of corporate bonds, insurers have long played a role as major buyers in the primary market for corporate bonds. Recent research has focused on another important aspect of insurers—their funding structure (Koijen and Yogo 2023; Chodorow-Reich, Ghent, and Haddad 2021). Due to the long contractual nature of insurance policies, insurers have stable funding that is little affected by market disruptions. This gives them the potential to be the "value investors," willing to buy when a strong desire for market selling presents profit opportunities.

Did insurers actually play such a stabilizing role in the COVID-19 bond market crises? Drawing on a variety of data sources, O'Hara, Rapp, and Zhou (2023) assemble detailed information on all insurance company secondary market bond trades. They also identify the dealer counterparties to these trades, providing a way to determine who the insurer traded with and how this affected the dealers' behavior. Their analysis reveals that insurers played a major stabilizing role in the March 6 to March 19 period prior to the Fed's involvement. While dealer bond inventories fell by a net $5 billion over this period, insurers were net buyers of $2.5 billion in corporate bonds. Insurers with greater stable funding were the most likely to be net buyers over this stress period. In addition, insurers bought more bonds from dealers with whom they had relationships, and they sold more to dealers without such relationships. Insurance company buying thus provided a backstop to dealers, who could transact with other clients knowing that they could potentially sell unwanted inventory positions to their insurance clients.

That long-term investors can play such a stabilizing role in the bond market is reassuring, but this role may be limited by regulatory and other constraints. Insurance companies are subject to capital requirements based on the riskiness of their asset holdings. While investment grade corporate bonds have relatively low capital charges, high-yield bonds face higher hurdle rates, suggesting that insurers' purchasing will be more focused on higher-rated bonds. Other long-term investors such as pension funds may also play a role, but may be hampered by less stability in their funding sources. What may constrain liquidity provision by any of these investors is uncertainty about what a bond is worth, particularly in a stressed market. As we discuss below, market structure issues may play an important role in addressing this problem.

Enhancing Transparency

Our discussion thus far has been about liquidity provision, but markets have an equally important role to play in price discovery. Price discovery refers to the process

by which new information is impounded into prices. A variety of factors enter into this process, including announcements on interest rate changes, company-specific news, and prices from related instruments such as credit default swaps, but by far the most important factors are the price where a bond recently traded and the trading prices of closely related bonds. How efficiently markets impound such information into prices is clearly important to any investor, but so, too, is access to this information. Without transparency into the price process, market stability is hard to maintain.

In US equity markets, transparency into the price process is provided at two levels. Investors want to know both a stock's most recent trading price and also the current quotes, or the prices at which they can expect to trade. This information is provided by the consolidated tape, a real-time information source set up as part of the National Market System, established by the Securities Acts Amendments of 1975 and now cosponsored by the National Association of Securities Dealers and the NASDAQ Stock Exchange. The tape, along with various reporting mandates, links the disparate US equity markets into one consolidated market.

The corporate bond market has more limited transparency. In 2005, the Financial Industry Regulatory Authority (FINRA) created the Trade Reporting and Compliance Engine as a reporting facility for over-the-counter fixed income trades. Because of the decentralized nature of corporate bond trading, TRACE data provide much-needed transparency into traded bond prices. But unlike equity prices, which are reported with millisecond delays to the consolidated tape, bond trades can take up to 15 minutes from execution to reporting. This delay can be consequential in volatile market conditions.

Moreover, FINRA neither collects nor disseminates pre-trade transparency data for corporate bonds. There are a variety of privately available pre-trade propriety data tools, the best known of which are from providers such as Bloomberg, MarketAxess, Neptune, and Algomi. While these pricing models are widely used by many institutional investors, each provider generally has their own proprietary model, so issues of uniformity arise. Proposals to add quote information for the most actively trading corporate bond issues and to disseminate these data publicly have been put forth, but as yet that has not come to pass. Pre-trade transparency may be particularly important in times of market stress. As prices become more uncertain, nondealer participants may be unwilling to post quotes on all-to-all electronic platforms, or, in the case of long-run investors, even step in to provide liquidity. Recently, regulators have turned attention to transparency issues in the market for Treasuries, with the goal of increasing the resiliency of that market. Enhancing market transparency in the corporate bond market may be an equally important area to consider.

Conclusions

The corporate bond market has changed. Market evolution has made trading cheaper, more competitive, and more technologically savvy. New investment

products have brought new investors to the market, and electronic trading has facilitated algorithmic and other trading innovations long employed by institutional investors in other asset classes. Viewed from an historical perspective, the bond market is clearly "better," not just for actively traded issues but for inactively traded issues as well.

What has not changed is that bond illiquidity is still an issue. The shift from a relationship-based to a transaction-based market, the greater interconnectedness of markets, the rise of electronic trading—all contribute to a market that can falter in stress times. In some ways, this is not new: dealers were always reluctant to "catch a falling knife" by buying when markets were plunging. What is new is the size, scale, and interconnectedness of the bond market, which amplifies the impact of any market instability. Moreover, the new reliance on bond funding for major banks highlights the systemic effects that can arise from bond market instability. In this new "mostly better" bond market, the Fed's recent role of market maker of last resort may prove to be more than a one-off event.

References

Adrian, Tobias, Nina Boyarchenko, and Or Shachar. 2017. "Dealer Balance Sheets and Bond Liquidity Provision." *Journal of Monetary Economics* 89: 92–109.

Andersen, Leif, Darrell Duffie, and Yang Song. 2019. "Funding Value Adjustments." *Journal of Finance* 74 (1): 145–92.

Bao, Jack, Maureen O'Hara, and Xing (Alex) Zhou. 2018. "The Volcker Rule and Corporate Bond Market Making in Times of Stress." *Journal of Financial Economics* 130 (1): 95–113.

Berndt, Antje, Darrell Duffie, and Yichao Zhu. 2025. "The Decline of Too Big to Fail." *American Economic Review* 115 (3): 945–74.

Bessembinder, Hendrik, Stacey Jacobsen, William Maxwell, and Kumar Venkataraman. 2018. "Capital Commitment and Illiquidity in Corporate Bonds." *Journal of Finance* 73 (4): 1615–61.

Bochner, Jacob, Min Wei, and Jie Yang. 2020. "What Drove Recent Trends in Corporate Bonds and Loans Usage?" FED Notes, October 23. https://doi.org/10.17016/2380-7172.2789.

Chodorow-Reich, Gabriel, Andra Ghent, and Valentin Haddad. 2021. "Asset Insulators." *Review of Financial Studies* 34 (3): 1509–39.

Choi, Jaewon, Saeid Hoseinzade, Sean Seunghun Shin, and Hassan Tehranian. 2020. "Corporate Bond Mutual Funds and Asset Fire Sales." *Journal of Financial Economics* 138 (2): 432–57.

Choi, Jaewon, Yesol Huh, and Sean Seunghun Shin. 2024. "Customer Liquidity Provision: Implications for Corporate Bond Transaction Costs." *Management Science* 70 (1): 187–206.

Dannhauser, Caitlin D. 2017. "The Impact of Innovation: Evidence from Corporate Bond Exchange-Traded Funds (ETFs)." *Journal of Financial Economics* 125 (3): 537–60.

Dick-Nielsen, Jens, and Marco Rossi, 2019. "The Cost of Immediacy for Corporate Bonds." *Review of Financial Studies* 32 (1): 1–41.

Duffie, Darrell. 2012. "Market Making under the Proposed Volcker Rule." Rock Center for Corporate Governance at Stanford University Working Paper 106.

Duffie, Darrell, Nicolae Gârleneau, and Lasse Heje Pedersen. 2005. "Over-the-Counter Markets." *Econometrica* 73 (6): 1815–47.

Finnerty, John D., Natalia Reisel, and Xun Zhong. 2024. "ETFs, Creation and Redemption Processes, and Bond Liquidity." *Journal of Financial and Quantitative Analysis*. https://doi.org/10.1017/S0022109024000346.

Falato, Antonio, Itay Goldstein, and Ali Hortaçsu. 2021. "Financial Fragility in the COVID-19 Crisis: The Case of Investment Funds in Corporate Bond Markets." *Journal of Monetary Economics* 123: 35–52.

Goldstein, Itay, Hao Jiang, and David T. Ng. 2017. "Investor Flows and Fragility in Corporate Bond Funds." *Journal of Financial Economics* 126 (3): 592–613.

Goldstein, Michael A., and Edith S. Hotchkiss. 2020. "Providing Liquidity in an Illiquid Market: Dealer Behavior in US Corporate Bonds." *Journal of Financial Economics* 135 (1): 16–40.

Haddad, Valentin, Alan Moreira, and Tyler Muir. 2021. "When Selling Becomes Viral: Disruptions in Debt Markets in the COVID- 19 Crisis and the Fed's Response." *Review of Financial Studies* 34 (11): 5309–51.

Hendershott, Terrence, Dan Li, Dmitry Livdan, and Norman Schürhoff. 2020. "Relationship Trading in Over- the-Counter Markets." *Journal of Finance* 75 (2): 683–733.

Hendershott, Terrence, Dmitry Livdan, and Norman Schürhoff. 2021. "All-to-All Liquidity in Corporate Bonds." Swiss Finance Institute Research Paper 21–43.

Hendershott, Terrence, Ananth Madhavan. 2015. "Click or call? Auction versus Search in the Over-the-Counter Market." *Journal of Finance* 70 (1): 419–47.

Holden, Craig W., and Jayoung Nam. 2024. "Market Accessibility, Bond ETFs and Liquidity." *Review of Finance* 28 (5): 1725–58.

Johnson, Steve. 2023. "Bond ETFs Suck Liquidity Out of Market in a Crisis, Academics Say." *Financial Times*, February 19. https://www.ft.com/content/d13d2c2f-0411-42ea-94dd-42331be05f9a.

Koijen, Ralph S. J., and Motohiro Yogo. 2023. "Understanding the Ownership Structure of Corporate Bonds." *American Economic Review: Insights* 5 (1): 73–92.

Kozora, Matthew, Bruce Marshall Mizrach, Matthew Peppe, Or Shachar, and Jonathan Sokobin. 2020. "Alternative Trading Systems in the Corporate Bond Market." Federal Reserve Bank of New York Staff Report 938.

Koont, Naz, Yiming Ma, Ľuboš Pástor, and Yao Zeng. 2023. "Steering a Ship in Illiquid Waters: Active Management of Passive Funds." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4053844.

Li, D., and N. Schürhoff. 2019. "Dealer networks." *Journal of Finance* 74(1): 91–144.

Li, Yi, Maureen O'Hara, and Xing (Alex) Zhou. 2024. "Mutual Fund Fragility, Dealer Liquidity Provisions, and the Pricing of Municipal Bonds." *Management Science* 70 (7): 4802–23.

Ma, Yiming, Kairong Xiao, and Yao Zeng. 2022. "Mutual Fund Liquidity Transformation and Reverse Flight to Liquidity." *Review of Financial Studies* 35 (10): 4674–4711.

Macchiavelli, Marco, and Xing (Alex) Zhou. 2022. "Funding Liquidity and Market Liquidity: The Broker-Dealer Perspective." *Management Science* 68 (5): 3379–98.

O'Hara, Maureen, Andreas C. Raap, and Xing (Alex) Zhou. 2023. "Bond Market Resiliency: The Role of Insurers." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4151934.

O'Hara, Maureen, Yihui Wang, and Xing (Alex) Zhou. 2018. "The Execution Quality of Corporate Bonds." *Journal of Financial Economics* 130 (2): 308–26.

O'Hara, Maureen, and Xing (Alex) Zhou. 2021a. "The Electronic Evolution of Corporate Bond Dealers." *Journal of Financial Economics* 140 (2): 368–90.

O'Hara, Maureen, and Xing (Alex) Zhou. 2021b. "Anatomy of a Liquidity Crisis: Corporate Bonds in the COVID-19 Crisis." *Journal of Financial Economics* 142 (1): 46–68.

O'Hara, Maureen, and Xing (Alex) Zhou. 2022. "Corporate Bond Trading: Finding the Customers' Yachts." *Journal of Portfolio Management* 48 (6): 96–109.

Pan, Kevin, and Yao Zeng. 2019. "ETF Arbitrage under Liquidity Mismatch." European Systematic Risk Board Working Paper 59.

Schultz, Paul. 2017. "Inventory Management by Corporate Bond Dealers." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.2966919.

Sibert, Anna, and Willem Buiter. 2007. "The Central Bank as the Market Maker of last Resort: From lender of last resort to market maker of last resort." *VoxEU*, August 13. https://cepr.org/voxeu/columns/central-bank-market-maker-last-resort-lender-last-resort-market-maker-last-resort?utm_source=chatgpt.com.

Todorov, Karamfil. 2021. "The Anatomy of Bond ETF Arbitrage." *BIS Quarterly Review*: 41–53.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 235–260

Why Is the Fragmented Municipal Bond Market So Costly to Investors and Issuers?

John M. Griffin, Nicholas Hirschey, and Samuel Kruger

The municipal bond market consists of over $4 trillion in outstanding debt (Federal Reserve 2025), which is utilized by cities, counties, and states to finance roads, schools, sewers, and other public investments. Municipal bonds also play a significant role in the investment portfolios of many households. Yet the municipal market suffers from multiple inefficiencies such as high and variable trading costs, high underwriting costs, portfolio holdings that are not tax-efficient, and use of costly credit ratings and insurance despite limited benefits. These frictions and costs make the market more expensive for investors and ultimately increase borrowing costs for municipalities, which is costly for taxpayers and leads to reduced provision of municipal services. This article offers a tour of the puzzling and variable nature of this market and what might be done to reduce the cost of state and local government financing.

One prominent inefficiency is the high cost of trading municipal bonds. In the 1920s, municipal bonds had trading costs similar to stocks. Today, retail investors pay 30 times more to trade municipal bonds than stocks, despite stocks being much more volatile and informationally sensitive. This difference results from a dramatic decline in stock trading costs, seemingly due to equities having more successful

■ *John M. Griffin is Professor of Finance, and Samuel Kruger is Associate Professor of Finance, both at the McCombs School of Business, University of Texas at Austin, Austin, Texas. Nicholas Hirschey is Assistant Professor of Finance, Nova School of Business and Economics, Universidade NOVA de Lisboa, Carcavelos, Portugal. Their email addresses are john.griffin@utexas.edu, nicholas.hirschey@novasbe.pt, and sam.kruger@mccombs.utexas.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241411. One of the authors is an officer of an organization whose interests relate to the article and has received financial support from that organization.

technological and regulatory innovations. Current trading costs in municipal bonds are highly variable and often swallow more than one-tenth of an investor's return in a seemingly safe asset. These inefficiencies are troubling given the size of the market.

Relative to bonds issued by the federal government and corporations, municipal bonds are noteworthy for the sheer number and varied financial sophistication of their issuers, which include approximately 50,000 cities, counties, states, and other government entities, many of which lack the financial sophistication and profit incentives of corporate bond issuers. We draw on research to explain how this market functions, highlighting inefficiencies and potential conflicts of interest, including issues related to underwriting, credit ratings, insurance, and tax treatment. We then propose changes to the market that could result in significant savings for municipalities and investors, focusing on improvements in the availability of information about prices and competitiveness of price quotes for potential buyers; increased transparency about municipalities' issuance expenses, standardized accounting statements, and potential conflicts of interest; and improved investment offerings for households, especially from mutual funds and exchange-traded funds. Because market participants earn significant fees from the current market structure, regulatory intervention is likely required to facilitate a more competitive environment.

Municipal Bond Market Background

Here we briefly describe the issuance process, trading features, and regulatory structure of the market. Our survey is not intended to be exhaustive.[1]

Market Structure

The largest use of municipal debt is for education, followed by utilities and transportation (IRS 2025). The bonds funding these investments are classified into two primary types. Revenue bonds are tied to income from specific assets such as public utilities, toll roads, stadiums, and other revenue-generating infrastructure. By contrast, general obligation bonds are not tied to a specific asset. Instead, they are backed by the general creditworthiness and taxing authority of the issuer. Certain projects favor one type of bond over the other. Debt that funds education is about three times more likely to be general obligation, whereas utility and transportation projects slightly favor revenue bonds. Both types of bonds are routinely issued by states, cities, counties, and other governmental entities.

When a municipality issues a bond, it borrows money from investors in exchange for a promise to pay a bond's owner a face value at the end of the bond's

[1] Readers interested in additional detail might begin with Cestau et al. (2019b), who discuss municipal bond history, the municipal bond puzzle, liquidity, and green municipal bonds; Bergstresser (2023), who surveys the municipal bond literature on a variety of issues including the history, real impacts, and regulatory frameworks; and Schleicher (2023), who evaluates options for managing defaults on municipal debt with extensive historical background on the municipal bond market.

life (maturity) and regular fixed coupon payments between issuance and maturity. Coupons function as interest payments on the bond and are quoted as a coupon rate, which is the annual coupon payment as a percentage of face value. Coupon interest payments on municipal bonds are generally exempt from federal income tax and are also typically exempt from state income taxes if the bond is issued in the same state where the investor resides. These tax exemptions function as subsidies for debt-financed municipal spending. In states with high income taxes, in-state tax exemptions incentivize high-income investors to purchase muni bond debt issued within their home state (Babina et al. 2021).

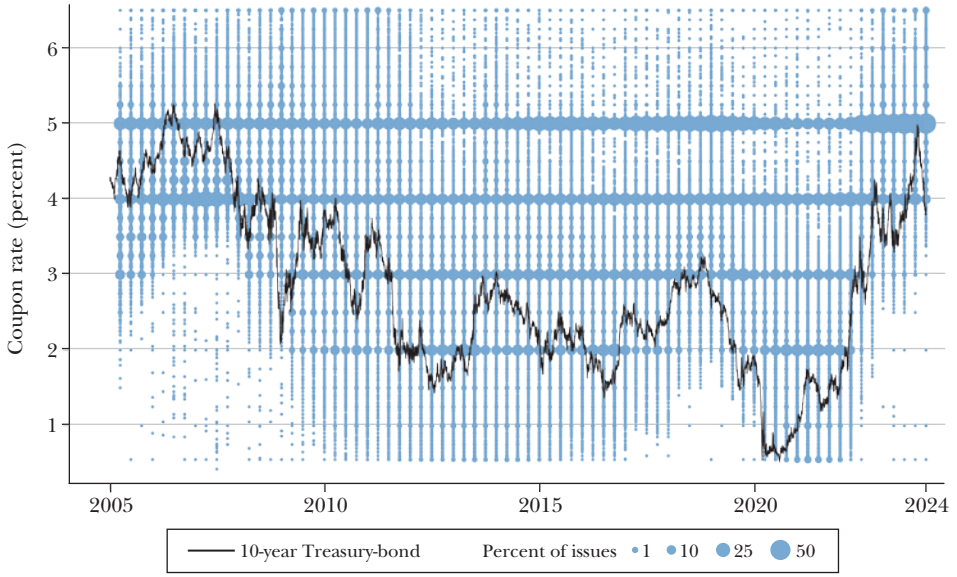
In addition to coupon payments, bond investors also earn a capital return on the difference between a bond's purchase price and the selling price (or face value, if held until maturity). A bond's price can differ from its face value even at issuance, and bond prices after issuance move in response to market interest rate changes as well as news about credit risk. Bonds priced above face value are referred to as trading "at a premium," and bonds with prices below their face value are trading "at a discount."

Despite their interest being tax-exempt, capital appreciation on municipal bonds is generally taxable (Ang, Bhansali, and Xing 2010a; Bagley et al. 2023). This tax treatment incentivizes issuers to issue bonds at a premium to reduce the chance that a bond will have a taxable capital gain. About 89 percent of municipal bonds are issued at a premium by offering elevated coupon payments that are offset by the issuer receiving a higher market price when selling the bonds (Landoni 2018).

Figure 1 shows the percentage of municipal bonds issued each quarter from January 2005 to December 2023 at different coupon rates; the *x*-axis shows the quarter of issuance and the *y*-axis shows the coupon rate. Larger circles indicate higher percentages of that quarter's bonds issued at a given coupon rate. Coupon interest rates typically exceed the 10-year Treasury bond yield and cluster at round numbers such as 3, 4, or 5 percent. From 2005 to 2007, 10-year Treasury bond yields were above 4 percent. During this time period, 23 percent of municipal bond issues had 4 percent coupons, 20 percent of municipal bonds had 5 percent coupons, and coupon rates below 3 percent were rare in order to ensure that bonds would be issued at a premium. When interest rates fell in the 2010s, 4 percent and 5 percent coupon rates remained common, and it also became possible to issue bonds at lower coupon rates. For example, from 2012 to 2021, the 10-year Treasury bond yield was nearly always below 3 percent, and an average of 11 percent of municipal bonds were issued at 2 percent coupon rates. When interest rates rose in 2022, coupon rates below 3 percent once again disappeared. By 2023, over half of all municipal bonds were issued with 5 percent coupon rates.

The initial issuance of municipal bonds (that is, the primary market) involves issuers selling bonds to investors, using underwriters as intermediaries. With more than 50,000 municipal bond issuers and approximately one million municipal securities (Bergstresser 2023), most issuers are not large, sophisticated players, and many issuers need help to navigate the market. In approximately 70 percent of cases, issuers employ a "municipal advisor" to help select an underwriter and determine

Figure 1
Coupon Rate Clusters



Source: We include all municipal bonds with positive coupon rates in the transaction reports and use the bond's dated date, the date it begins to accrue interest, to assign bonds to issuance quarters. Coupon rates are winsorized at 0.5 percent and 99.5 percent levels. Bonds are from the Municipal Securities Rulemaking Board, Treasury yields from the Federal Reserve's H.15 release (series DGS10).
Note: This figure shows the percentage of municipal bonds issued at particular coupon rates each quarter from January 2005 to December 2023. The size of the blue circles shows the percentage of bonds issued at that coupon rate that quarter. The black line shows the ten-year constant maturity Treasury yield.

the bond's structure, sale, and derivative use (Luby and Hildreth 2014). Issuers may also insure the bond's cash flows against default, hoping to lower the interest rate the municipality pays. Issuers typically also pay a credit rating agency to issue a rating on the bonds, and investors use this rating to help understand the bond's risk.

After municipal bonds are issued, they trade in an "over-the-counter" secondary market, in which customers buying or selling a bond trade with dealers who source bonds from either their own inventory or a network of other dealers called the "interdealer market." Because municipal bonds are often held for long periods, trading volume is highest for newly issued bonds (Green, Hollifield, and Schürhoff 2007a). Electronic trading platforms exist only in the interdealer market, with no widely used platform for individuals to trade directly with each other. Municipal bond transactions are publicly disclosed after they occur, but individual investors have limited pre-trade information, particularly with respect to quoted prices (Securities and Exchange Commission 2012; Biais and Green 2019).

Whereas most financial markets are dominated by institutional investors, households are the largest direct holder of municipal bonds. Households also invest in municipal bonds through mutual funds and exchange-traded funds, which

represent a growing share of mutual fund ownership (Bagley, Vieria, and Hamlin 2022; Adelino et al. 2023). Using data from the Federal Reserve, we calculate that households currently own 44.5 percent of municipal bonds, and ownership by mutual funds and exchange-traded funds has increased from 15.6 percent in 2004 to 24.2 percent in 2024. An increasing share of household holdings are held through separately managed accounts, in which wealth managers tailor portfolios to the household's circumstances and tend to be compensated by a fee that is a percentage of assets rather than by commissions on trades (Albright 2024). Muni-bond ownership is quite different from corporate bonds, where households hold less than 5 percent of the bonds.

Regulatory Framework

The municipal bond market is primarily overseen by the Municipal Securities Rulemaking Board (MSRB), a self-regulatory organization created in 1975 with the goal that the municipal bond market be "fair and efficient." Municipal bonds are exempt from reporting requirements to the Securities and Exchange Commission (SEC), but MSRB rules must be approved by the SEC, and the SEC has authority to bring fraud enforcement actions. Exemptions from SEC reporting requirements typically result in less standardized accounting data related to municipal bonds for everything from underwriting expenses to the fiscal conditions of municipalities. The Financial Industry Regulatory Authority (FINRA), along with the MSRB and the SEC, examines compliance of brokers, dealers, underwriters, and municipal advisors. Regulatory reforms over the past 20 years include tighter rules for potential conflicts of interest for municipal advisors, increased post-trade price transparency, and a requirement that dealers give their customers "best execution" on municipal bond trades. Evidence on the impact of these reforms is mixed, as we discuss in more detail below.

High and Inconsistent Trading Costs

In order to buy and sell municipal bonds in the secondary market, investors incur trading costs. These costs matter because they directly affect the net returns that investors experience in the market. To the extent that trading costs make the market less attractive to investors, they likely also necessitate higher interest rates in the primary market, which increases costs for municipalities and could negatively affect provision of municipal services.

Trading Costs

Historical experience shows it is possible for municipal bonds to have transaction costs similar to equities. During the 1920s, municipal bonds traded on the New York Stock Exchange along with equities, and they both had trading costs around 1 percent (Biais and Green 2019; Jones 2002). Subsequently, municipal bond trading moved off the exchange to the over-the-counter market, where it remains

today. Biais and Green (2019) argue the move was caused by demand for more physical space to trade equities and a shift towards more institutions and fewer individuals trading municipal bonds. Whatever the reason, while equities and municipal bonds had similar costs for retail investors 100 years ago, their costs are dramatically different today.

Indeed, one of the most fundamental questions for bond markets is why trading costs are significantly higher than in equity markets (Bessembinder, Spatt, and Venkataraman 2020). In addition, small municipal bond trades have significantly higher trading costs than large trades (Harris and Piwowar 2006; Green, Hollifield, and Schürhoff 2007a), which is potentially surprising because they involve less inventory risk for dealers and less adverse selection (because retail investors are typically less informed than institutions). In equity markets, by contrast, trading costs are typically lower for retail investors than for institutions.

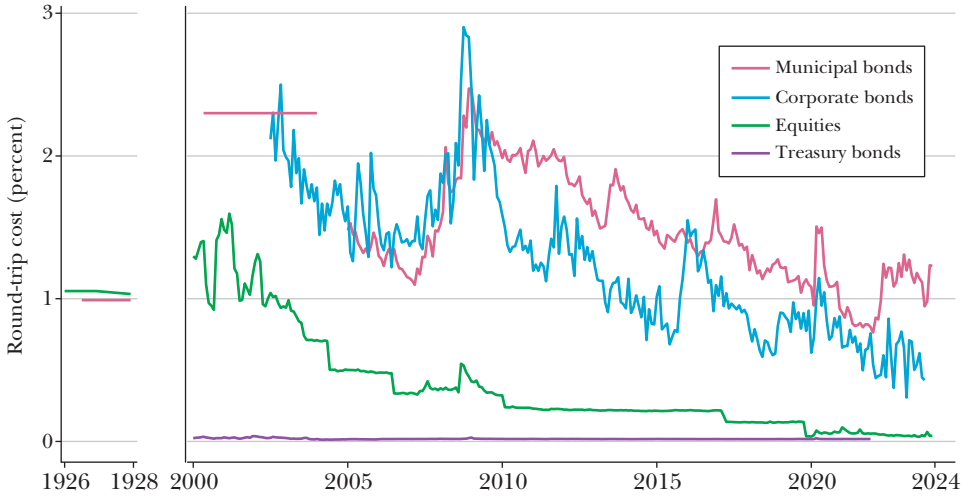
Figure 2 shows average "round-trip" trading costs for retail investors in municipal bonds compared to equities, corporate bonds, and Treasury bonds from roughly 2000 to 2023. The round-trip cost is the cost to buy a bond plus the cost to sell it, including markups and brokerage commissions. This measure is relevant because if you buy a bond and sell it later, you earn the bond's return minus the round-trip cost. The data are a combination of our own estimates and estimates from other papers.

In 2000, trading costs for equities were not much different from back in the 1920s, but since then, these costs have fallen sharply from 1.06 percent to 0.04 percent. By contrast, retail investors' trading costs for municipal bonds were 2.3 percent in the early 2000s (Green, Hollifield, and Schürhoff 2007b), and while they have fallen in the last decade, they were still 1.1 percent in 2023. Remarkably, trading costs for retail municipal bond investors are higher now than they were in the 1920s. In addition to trading costs being high on average, extreme bond markups are common and have only experienced mild decreases over the past 30 years. In particular, we estimate that 5 percent of small bond transactions have markups over 3.29 percent in 2023. This is despite the long-standing Municipal Securities Rulemaking Board (MSRB) Rule G-30, which requires that commissions and markups for transactions with dealers be "fair and reasonable," and Rule G-17, which states each dealer and advisor must "deal fairly with all persons." The fact that trading costs for municipal bonds are higher than equities and did not fall nearly as much from 2000 to 2023 is consistent with the contention that trading costs are high at least in part because the municipal bond market lacks a trading platform that individuals can access (Harris, Kyle, and Sirri 2015; Bessembinder, Spatt, and Venkataraman 2020).

Municipal bond trading costs are also high compared to other bonds. Treasury trading costs are extremely low throughout our sample period beginning in 2000. Corporate bond trading costs were similar to municipal bonds from 2002 to 2009, but fell more rapidly than municipal bond costs and are approximately half of municipal bond trading costs at the end of 2023.

The Municipal Securities Rulemaking Board made changes to make the municipal bond market more transparent and efficient, but with limited success. Implementation of the MSRB Real-Time Transaction Reporting System (RTRS) in

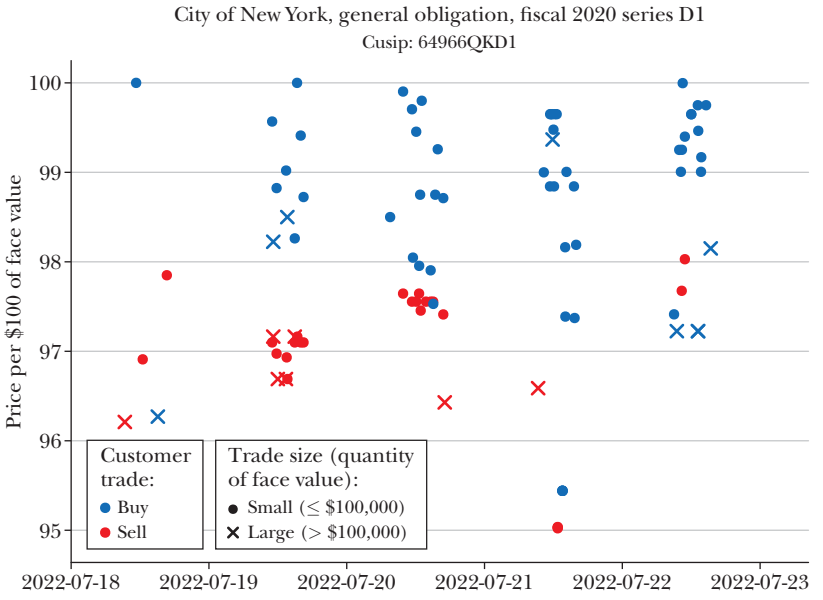
Figure 2
Round-Trip Costs for Small Trades



Source: Municipal bond transaction costs from July 1926 to December 1927 are from Biais and Green (2019), from May 2000 to January 2004 they are for purchases less than $100,000 from Green, Hollifield, and Schürhoff (2007b), and from January 2005 to December 2023 they are calculated by the authors for purchases less than or equal to $100,000 from Municipal Securities Rulemaking Board (MSRB) trade reports. Equity costs from January 1926 to December 1927 are for NYSE stocks from Jones (2002), and from January 2000 to December 2023 they are calculated by the authors for NYSE, NASDAQ, and AMEX stocks using CRSP data and Schwab brokerage commissions. Corporate bond transaction costs from July 2002 to September 2023 are calculated by the authors from Trade Reporting and Compliance Engine (TRACE) trade reports. Treasury bond costs from January 2000 to December 2021 are from Adrian, Fleming, and Vogt (2023). Griffin, Hirschey, and Kruger (2025) provide additional details.
Note: This figure shows round-trip transaction cost estimates for small trades representative of retail traders. Municipal bond costs from the 1920s are the proportional quoted spread plus round-trip commission. Municipal and Corporate bond costs since 2000 are the percent premium of a dealer's sale price to customers over their purchase price from customers. Equity costs are the proportional bid-ask spread plus round-trip commission. Treasury costs are the proportional bid-ask spread.

2005, which reports trade information within 15 minutes instead of a one-day delay, led to a decrease in trading costs in the secondary market (Chalmers, Liu, and Wang 2021), but had little effect on average markups in the primary market (Schultz 2012). Neither the MSRB's "best execution" Rule G-18 implemented in 2016, requiring dealers to obtain the most favorable possible price, nor the required disclosure of markups and markdowns in 2018 (MSRB Regulatory Notice 2016-28) had a discernible effect on average or extreme markups (Griffin, Hirschey, and Kruger 2023). Along the same lines, in 2019 the Securities and Exchange Commission (SEC) adopted Regulation Best Interest, which codifies a broker's responsibility to pursue retail customers' best interests. The trading cost trends in Figure 2 show that trading cost decreases for municipal bonds have been slow, particularly compared to other bond markets.

Figure 3
Example of Large Price Variation



Source: Municipal Securities Rulemaking Board trade reports.
Note: This figure shows a week of trading for a bond with large variation in customer transaction prices. Small trades have less than or equal to $100,000 face value, Large trades have more than $100,000 face value. Interdealer trades and trades from customers who do not pay transaction-based fees are excluded.

Extreme and Inconsistent Costs

The markups on municipal bonds during trading are frequently extreme and highly variable, particularly for retail investors. Figure 3 shows an extreme example of price variation on large and small trades on a specific New York City municipal bond, trading in July 2022. The price on the vertical axis is an amount per $100 of face value. Trade size is the total quantity of face value in the transaction. Trades of less than or equal to $100,000 of face value, which we call "small trades" and indicate with circles in the figure, are more likely between a retail customer and a dealer. Trades of more than $100,000 of face value, indicated with crosses, are more likely between an institution and a dealer. First, small buy transactions have a wide range of prices. Throughout July 19 to July 22, some small purchases paid as much as $100 per $100 face value of the bond, while other small purchases paid $98 or less. Prices for large trades are less extreme, with only one trade being executed above $98.50.

Second, customer sell transactions are less volatile, but also considerably lower than the buy transactions. As a result, dealers are making spreads ranging from 1 to 3 percent. Because municipal bonds typically do not experience significant intra-day news events or large intra-day interest rates changes, this variation cannot be due to changing fundamentals. This type of variation is far from unique. In fact, intraday

price dispersion of at least 1 percent occurs 44 percent of the time for small trades of the same bond on the same day and substantial variation even occurs within the same dealer's trades on the same day (Griffin, Hirschey, and Kruger 2023).

Such variation in pricing for the same bond on the same day seems to contradict guidance for Municipal Securities Rulemaking Board Rules G-30 and G-18, which state that trades on the same day should generally have the same price. Additionally, the Securities and Exchange Commission has previously enforced such actions. In particular, Edward Jones was fined by the SEC based on allegations that the firm had violated Rule G-30 by selling bonds in some trades at higher prices and lower yields than other trades received, even for exactly the same bond on the same day (for the cease-and-desist order, see Securities and Exchange Commission 2015).

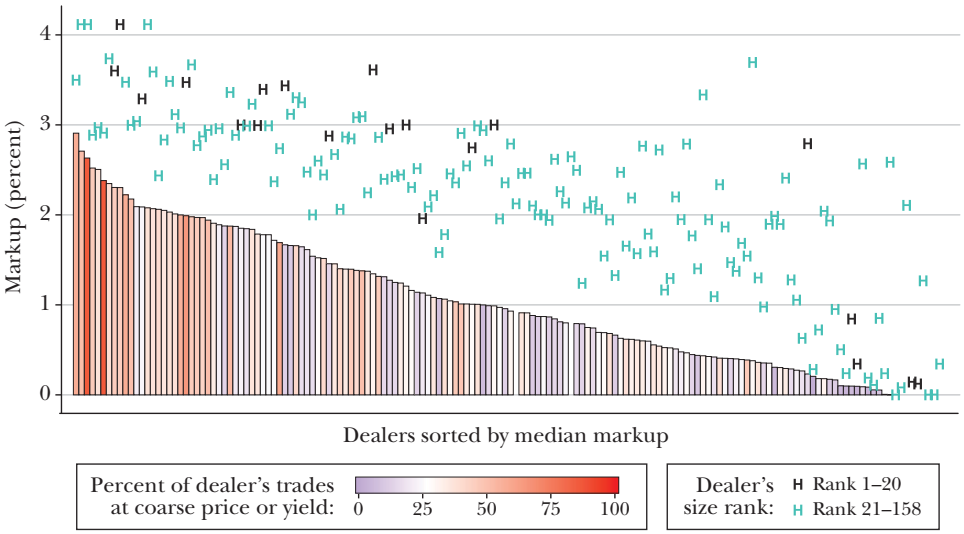
Why Do Trading Costs Vary across Dealers?

The research literature proposes several nonmutually exclusive explanations for why municipal bond prices vary so widely, even for the same bond on the same day: lack of trade transparency (Harris and Piwowar 2006; Harris 2015), costs from intermediating transactions through multiple dealers (Schultz 2012), differential dealer costs due to network centrality (Li and Schürhoff 2019), and the strategic use of market power (Green 2007; Green, Hollifield, and Schürhoff 2007b). For example, Green, Li, and Schürhoff (2010) find that when bond yields change, dealers raise prices quickly but delay lowering them, which indicates that dealers exercise market power resulting from decentralization and lack of price transparency in the municipal bond market.

Municipal bond trading costs vary widely across dealers. In Figure 4, we calculate markups on small customer purchases less than or equal to $100,000 by the largest municipal bond dealers by trade volume, following the methodology used by Griffin, Hirschey, and Kruger (2023). The anonymized dealer-level data are from July 2011 to December 2017. Of the 158 dealers, 25 dealers have median markups below 25 basis points and 11 have median markups below 5 basis points. However, 19 dealers have median markups above 2 percent, and 74 dealers (which represents 47 percent of all dealers) have markups above 1 percent. We show that large dispersions in prices are similar after controlling for bond characteristics and bond fixed effects, even when examining trades of the same bond on the same day. These findings indicate the differences across dealers are not driven by differences in the types of bonds that the dealers sell.

Rounding off prices may be a convenient heuristic to increase spreads, as found by Christie and Schultz (1994, 1995) in equities 30 years ago. The colors of the bars in Figure 4 show that dealers with high costs to the left of the figure are frequently in dark red, indicating they are more likely to use rounded prices or rounded yields. Dealers with costs below 25 basis points rarely trade with rounded prices or yields (for similar evidence from an earlier study, see Li 2007). If prices simply reflect dealer costs, there is little economic reason for prices to be rounded. Consistent with this intuition, Griffin, Hirschey, and Kruger (2023) find that cost-based explanations such as dealer network centrality and liquidity explain almost none

Figure 4
Variation in Small-Trade Markups across Dealers



Source: Transactions from the Municipal Securities Rulemaking Board's academic dataset with anonymized dealer IDs. The markup is customers' estimated round-trip cost following Griffin, Hirschey, and Kruger (2023). Griffin, Hirschey, and Kruger (2025) provide additional details.
Note: This figure shows the median and 95th percentile of dealers' seasoned-issue markups in small (less than or equal to $100,000) sales to customers from July 2011 to December 2017. We limit the figure to dealers who have at least $20 million worth of small trades with customers. The bar height is the median markup, and color shows the percentage of a dealer's trades that are at coarse prices or coarse yields. Transactions at coarse prices or coarse yields are those where the fractional part of the price or yield is at exact quarters or odd eighths, such as a price of $100.50 or a yield of 3.375 percent. The "H" above the bar is the 95th percentile, and color shows the dealer's size rank.

of the observed variation in markups across dealers. Instead, they suggest the possibility of a segmented market in which some dealers focus on extracting rents from their customers through strategic behavior, while other dealers prioritize low costs.

Dealers also vary in their propensity to employ extreme markups, and extreme bond markups cannot simply be blamed on small dealers. Figure 4 also plots the 95th percentile of markups for all dealers. The top 20 dealers by size are plotted in black, and dealers outside the top 20 are plotted in light green. First, markups around 3 percent or more are relatively common across most dealers. Second, among the 20 biggest dealers, a large number have markups at or above 3 percent; indeed, large dealers have among the highest incidence of extreme markups. Third, there are at least two large dealers whose 95th percentile markups are nearly zero, indicating that consistently low pricing is possible and sustainable for some dealers.

The large number of extreme markups is striking and appears to have generated little regulatory or disciplinary scrutiny. For example, disciplinary actions are

rare even though Financial Industry Regulatory Authority (FINRA) has indicated that markups over 3 percent are potentially ripe for discipline (Griffin, Hirschey, and Kruger 2023). While it would be nice for market participants and researchers to be able to identify which dealers consistently exhibit low markups and which dealers have high and inconsistent markups, the dealers are anonymized in the data used for this analysis.

The marketing literature shows that firms often set prices to exploit customers' cognitive biases. If dealers are focused on extracting rents from customers, they might be apt to mark up bond prices as much as possible without crossing salient yield thresholds. For example, a customer might react similarly to a yield of 4.01 percent instead of 4.08 percent while perceiving a yield of 3.99 percent to be noticeably lower. Griffin, Hirschey, and Kruger (2023) show strong discontinuities around percentage point yields, with a higher prevalence of yields just above these thresholds and higher markups on these trades compared to trades just below round number yield thresholds. Dealers appear to use their price discretion to raise markups and decrease yields without crossing salient thresholds that customers would notice. This practice is also much more common at high-cost dealers than those with low trading costs.

Conflicts of Interest in the Primary Market

Municipal bond issuance is a complicated transaction with significant information asymmetries and possible conflicts of interest between the issuer, municipal bond advisor, underwriter, and credit rating agencies. To the extent that these conflicts result in extra fees or other inefficiencies, this directly affects municipal borrowing costs and could also impede the provision of municipal services. We organize our discussion around the major players in the market: underwriters, municipal advisors, issuers, credit raters, and insurance.

Underwriters

Underwriters function as intermediaries between issuers and investors in the municipal debt market, playing a certification and operational role in bringing securities to market. Underwriting is a function often performed by large investment banks, including Barclays, Citigroup, Goldman Sachs, Jefferies, J. P. Morgan, Merrill Lynch, Morgan Stanley, and Wells Fargo. National underwriters make up about half of underwriting volume, with large and small regional underwriters filling the remaining share (Hund et al. 2024). Underwriting can be organized as either a negotiated or competitive offering. In a negotiated offering, a single underwriter helps the municipality structure the bond issue. In a competitive offering, the municipality first structures the bond issue (potentially with the assistance of a municipal advisor) and then selects an underwriter following an auction in which underwriters compete by specifying the yield at which they are willing to purchase the bonds.

The majority (57 percent) of new municipal bond issues are sold in competitive offerings, and state law requires many bond issues (depending on characteristics of the issue) to use competitive offerings (Cestau 2019). By contrast, 72 percent of refunding issues, which refinance previously-issued bonds, are negotiated. Cestau et al. (2019a) find that negotiated sales result in yields that are 15–17 basis points higher than competitive offerings, while state restrictions requiring competitive offerings reduce yields by 13 basis points. Nonetheless, 80 percent of issuers choose negotiated offerings when they are permitted to do so. Brancaccio and Kang (2022) describe how underwriters are incentivized to include special provisions like floating rates and special redemption, because these complexities give them a strategic advantage in secondary market trading. Cestau et al. (2019a) estimate that a nationwide requirement that issuers use competitive offerings would save issuers approximately $360 million annually. Several studies indicate that the underwriting market for municipal bonds is not fully competitive (Cestau 2019; Garrett et al. 2023).

Two major sources of underwriting profit are "underwriting spreads" and markups. The gross underwriting spread is the difference between the price at which the bonds are purchased from the municipality (the "takedown price") and the price at which the bond is resold to investors (the "reoffering price"). Underwriters can also profit by selling bonds at additional markups (Schultz 2012; Griffin, Hirschey, and Kruger 2023), similar to the secondary market markups discussed in the previous section. Hund et al. (2024) show average underwriting spreads in 2023 varying between around 65 basis points for national underwriters, slightly over 100 for regional underwriters, and over 200 basis points for single-state underwriters. They find that reported bond markups over the offering price have declined over time from 15 to 20 basis points in 2005–2017 to 5 basis points in 2023.[2]

Municipal underwriting presents numerous potential conflicts of interest with the possibility of additional hidden profits; for example, placing bonds with related parties at a discounted price, selling bonds with derivative features that favor the underwriter, favoring callable bonds for higher fees, or offering bonds with varying features such as longer maturity (which is associated with higher markups) that might generate more revenue for the underwriter. Because of the complexity in evaluating bond prospectuses, underwriter markups in primary sales, and the potential for hidden fees, it is difficult for either municipalities or academic researchers to evaluate these issues.

One infamous example of agency costs run amok happened in Jefferson County, Alabama, where employees of J. P. Morgan allegedly bribed local officials to secure underwriting business that charged Jefferson County extra fees on swap transactions on $4.3 billion in bonds dating back to 2003. Losses on these positions

[2] Including markups, Garrett (2024) finds average underwriting spreads of 57 basis points. The average markup included in this calculation combines the low markups to institutional traders with high markups to individuals. A more granular breakdown of major expense categories (for example, by Raineri et al. 2012) can include underwriter fees, underwriter expenses, management fees, and the profits from underwriter markups.

caused the county to declare bankruptcy in November 2011. Between fines paid and loan forgiveness, J. P. Morgan eventually paid $1.57 billion, and 17 county officials and contractors went to federal prison in connection with the financing and construction of a sewer system (Selway 2013; Faulk 2015).

It is unclear how widespread such problems might be. The Securities and Exchange Commission carried out a Municipalities Continuing Disclosure Cooperation initiative from 2014–2016 that encouraged underwriters to self-report misstatements in their offering documents in return for favorable terms. The initiative concluded with 72 underwriters, including most major names, settling and paying fines at or below $500,000 (Securities and Exchange Commission 2016). The settlements indicate the defendants willfully violated federal anti-fraud statutes by selling securities based on misleading disclosure statements, including false statements and omissions. Additionally, the 79 charges filed during the two-year period of the voluntary disclosure initiative greatly surpassed the ten cases in the prior two years, indicating undetected violations had probably been widespread (Carriel 2017).

One potential conflict of interest is that underwriters may be able to curry favor with connected institutional investors by offering them bonds at below-market prices. Indeed, Cao, Ye, and Wermers (2024) find that a significant source of profit for municipal bond funds is obtaining underpriced bonds at issuance. On average, they find that municipal bonds sold by a connected underwriter outperform other bonds by 14 basis points in the first month. To the extent that underwriters could have placed the bonds to other investors with lower yields without affecting underwriter spreads, this is a direct cost at the expense of issuers. Butler (2008) tests the role of local underwriters, who have more local information but might also be more apt to have political connections. They find that the local underwriters, on average, seem able to mitigate these political ties and issue bonds with lower fees and lower yields. However, Butler, Fauver, and Mortal (2009) find that states with more corruption have higher yields. Underwriter markups (gross spreads) are also higher in corrupt areas, but only during a period prior to 1994 when payments from underwriters to politicians were more feasible. This effect is economically large (12–14 basis points) and concentrated in negotiated bonds (rather than competitive offerings), where there is more opportunity for favoritism and conflicts of interest are more severe.

Municipal Advisors

Municipal advisors are generally separate firms that advise municipalities on issues such as how to structure bond offerings, how to invest bond proceeds, and the potential use of derivatives. Examples include Public Financial Management Inc., First Southwest Company Inc., and Ehlers and Associates. In general, their role is to function as independent advisors to municipalities, with the goal of reducing the information asymmetry between the investment bank (underwriter) and the municipality. Municipal advisors also face their own potential conflicts of interest; for example, municipal advisors could have other business interests that cause them to favor underwriters that are not in the municipality's best interest.

Based on data for eleven states obtained from Freedom of Information Act requests, Malakar (2024) finds that average municipal advisor fees fell from 40 basis points in 2010 to less than 20 basis points in 2021. Looking at data on municipal bond issuances and fees in the state of Texas, Luby and Moldogaziev (2013) find that issues with a municipal advisor have lower gross underwriter spreads, but underwriter spreads increase when there is more interaction between the advisor and the underwriter—which is consistent with conflicted advisors not acting in the municipalities' best interest. Because this study addresses only disclosed fees, it could understate conflicts of interest if agency issues are more severe for other hidden profit sources such as more opaque markups and derivative positions.

In the aftermath of the Wall Street Reform and Protection Act of 2010, better known as the Dodd-Frank Act, the authority of the Municipal Securities Rulemaking Board was expanded, particularly with respect to protecting municipalities. The MSRB amended its Rule G-23 to forbid financial advisors from jointly serving as underwriters, which affected approximately 15 percent of municipal bond issuances prior to the rule's implementation in November 2011. Industry reports like Bond Dealers of America (2019) argue that the rule change led to less competition. Yet Garrett (2024) finds that removing the conflicted advisors allows for more competition and that borrowing costs fell by about 5 percent for municipalities that previously had dual advisors. Another reform in June 2016 (Rule G-42) required that municipal bond advisors act as "fiduciaries" to their municipality clients, which includes a duty of care to be knowledgeable about the advice they give and a duty of loyalty requiring them to deal honestly and in good faith with their clients.

Refinancing

About 95 percent of long-term municipal bonds include a "call option," allowing the issuer of the bond to pay it off early at a preset price—for example, with the intention of refinancing at a lower interest rate (Chen, Cohen, and Liu 2024). But decisions about whether and when to refinance a bond can be complex, and issuers seem prone to costly mistakes.

For example, municipalities are able to refinance their debt even before their bonds are eligible to be called, by using advanced "refunding transactions" in which the municipality issues new debt to repay the old debt (subject to certain IRS rules). However, this option requires municipalities to precommit to calling their bonds, which is akin to early exercise of an option and typically destroys option value. A study of advance refundings found that 85 percent involve a loss of net present value (Ang et al. 2017). At the same time, municipal issuers also frequently wait too long before calling their bonds. By one estimate, delayed execution of call options costs municipal issuers $1.74 billion per year, and municipal advisors do not appear to alleviate delayed call execution (Chen, Cohen, and Liu 2024).

Credit Ratings

Municipal bonds respond to changes in credit ratings. In a 2010 episode, the credit rating agencies Moody's and Fitch modified their methodologies for

municipal credit ratings, leading to upgrades of zero to four credit rating notches on $2.2 trillion in municipal debt. This change in methodology did not alter actual underlying credit risk, so several studies seized upon this experience as a natural experiment for estimating the impact of credit ratings.

Comparing municipal bonds affected and unaffected by the change in bond ratings, Cornaggia, Cornaggia, and Israelsen (2018) find that the ratings upgrades led to an economically large 10 to 33 basis point decrease in credit spreads for the affected bonds—and the upgraded issuers responded by issuing more debt. The effects of bond credit ratings are more pronounced in settings where investors have less information such as issues with opaque accounting information, issues in corrupt states, and issues without alternative S&P credit ratings. Similarly, Adelino, Cunha, and Ferreira (2017) find that higher credit ratings allowed local governments, particularly those that are financially constrained, to expand their debt capacity and had positive economic multipliers during the recession. Gillette, Samuels, and Zhou (2020) find that firms with upgraded credit ratings decrease their financial disclosures; and Cheng, Cuny, and Xue (2023) find that relatively disadvantaged issuers increase the timeliness and quality of their accounting disclosures, indicating a tradeoff between ratings and financial disclosure.

Overall, the literature indicates that investors rely heavily on credit ratings, even when changes in ratings convey relatively little new information. As another example, Cornaggia, Hund, and Nguyen (2022) focus on how bond prices are affected by information on performance and solvency of the company insuring the bonds. They find that muni bonds did not price in the default risk of the insurers and only reacted after the insurers' credit ratings were actually downgraded.

The heavy reliance on credit ratings need not be problematic if the ratings are accurate. However, credit ratings are frequently biased and suffer from conflicts of interest due to issuers of bonds paying for the ratings. Compared to other markets, municipal bonds ratings are more straightforward because the bonds are simpler and less volatile (Cornaggia, Cornaggia, and Hund 2017). Nonetheless, rating errors in the municipal bond market may still be economically important. Using municipal bond data from Texas, where credit rating fees are disclosed, Cornaggia, Cornaggia, and Israelsen (2023) find that credit rating agencies tend to issue inflated municipal bond ratings (compared to their competitors) when they receive higher rating fees, and these inflated rates are less likely to be accurate in the future.[3] Adding to concerns over the issuer-pays model of ratings, Beatty et al. (2019) find that ratings upgrades associated with the 2010 recalibration by Moody's and Fitch discussed above also led to larger fees and increased market share for these rating agencies.

[3] Credit ratings that exhibit bias when paid for by the issuer of the financial instrument are pervasive in structured finance both during the financial crisis (Griffin and Tang 2012; Griffin 2021) and after the financial crisis (Baghai and Becker 2020; Griffin and Nickerson 2023). Similar problems are also prevalent in post-crisis corporate debt (Herpfer and Maturana 2021). Credit rating analysts also exhibit home bias and overestimate the creditworthiness of local issuers (Cornaggia, Cornaggia, and Israelsen 2023).

Do municipal credit ratings exhibit other errors? While credit rating agencies have historically advertised that credit ratings were similar across asset classes, it appears that municipal bonds are more likely to be upgraded than downgraded from 1980 to 2010, relative to the patterns of corporate bonds and structured finance (Cornaggia, Cornaggia, and Hund 2017). This finding suggests that rating agencies have historically penalized municipal bond issuers relative to other types of debt in a manner that could be costly to municipalities.

Municipal Bond Insurance

With reliable bond insurance, issuers can decrease the interest they need to pay on a bond, and the credit rating on the bond is essentially the credit rating of the insurance company. Before the financial crisis of 2008–2009, the gross value of insurance to bond issuers (in terms of lower interest payments) was similar to the insurance cost. However, after 2009, many insurance companies were no longer AAA-rated, and insurance often no longer provides sufficient value to bond issuers in terms of lower interest rates (Cornaggia, Hund, and Nguyen 2024). The fraction of municipal bonds that use insurance dropped from over 60 percent prior to the crisis to 20 percent or less from 2012 to 2020. For a sample of California and New York municipal bonds from 1996 to 2014, Bronshtein and Makridis (2020) find a sharply declining value of insurance, which they also attribute to the declining financial strength of insurers.

The current puzzle is why many municipal bond issuers who historically purchased insurance continue to do so. Habit may be the answer (Cornaggia, Hund, and Nguyen 2024). Of course, the benefits of bond insurance for issuers could change if the credit risk of municipal bond insurers improved or if the role of credit ratings were reduced in this market.

Would Less Retail Ownership of Municipal Bonds Be Preferable?

The complexity and potential for conflicts of interest in the municipal bond market are largely due to the fragmentation and relative lack of sophistication on both sides of the market—bonds issued by thousands of dispersed issuers and mainly held by millions of individual retail investors. The dispersed nature of bond issuance is inherent to the structure of government in the United States and is unlikely to change. Ownership of municipal bonds, on the other hand, could change. Would more ownership of municipal bonds by large sophisticated investors and less retail ownership be preferable to the current system?

Municipal Bond Tax Treatment

Tax exemptions for interest from municipal bonds create a strong incentive for them to be held by high-income federal taxpayers, who are willing to pay more for municipal bonds than other investors. State tax exemptions add an additional benefit for investing in bonds issued in the taxpayer's home state. These complicated

tax clienteles give rise to the potential for large mistakes. For example, a tax-exempt municipal bond is more valuable to someone with a 37 percent federal tax bracket than someone with a lower marginal tax rate. If enough investors in the top tax bracket buy the bond and push up its price, it will no longer be an attractive asset for investors with lower tax rates. Similarly, California bonds are more attractive to California investors and may not be as suitable for out-of-state investors. Determining what bonds are suitable for a particular investor after considering the effects of buying pressure from other investors with different tax incentives is a complicated decision with significant potential for investors to make mistakes.

A longstanding puzzle in the municipal bond literature is that short-term municipal yields are consistent with the tax rates we would expect for investors in the top income tax brackets, whereas longer-term municipal yields are higher than we would expect.[4] As a result, long-term municipal bonds appear to provide excess after-tax returns to investors in top tax brackets, indicating that municipal bond tax exemptions are a regressive transfer. However, even though the tax subsidies flow to high-income individuals, Garrett et al. (2023) find that the tax subsidy to municipal debt increases underwriter competition, which in turn leads to lower underwriting fees that make tax exemptions a cost-effective mechanism to lower borrowing rates.

If municipal bonds are priced such that the marginal investor is in one of the top tax brackets, any investors in lower tax brackets that choose to invest in municipal bonds are sacrificing significant after-tax returns compared to what they could earn from other investments. Bergstresser and Cohen (2016) find that the median marginal tax rate of households that own municipal bonds is 25 percent, and at least one-fourth of households that own municipal bonds have a zero percent marginal tax rate. Similarly, bonds in high-tax states are priced with lower yields because of their benefits to in-state taxpayers (Kidwell, Koch, and Stock 1984; Babina et al. 2021; Garrett et al. 2023). As a result, investors from other states should generally steer clear of these bonds. Common investing advice emphasizes the advantage of in-state bonds for residents of high tax states such as California and New York, but investing advice is sometimes less clear about the equally important implication that residents of other states should avoid these bonds because of their generally lower yields. For example, Howard (2024) advocates that residents in New York and California should invest in-state, whereas residents of other states should diversify nationally.

Mutual Funds and Exchange-Traded Funds

Federal Reserve data indicate that the share of municipal bonds owned by mutual funds and exchange-traded funds has risen from around 15 percent in the early 2000s to just below 25 percent at the start of 2024. Investors in these funds

[4] As a result, the municipal bond yield curve is steeper than the Treasury yield curve. Proposed explanations include limits to arbitrage, liquidity, and credit risk. Cestau et al. (2019b) summarize the academic literature on this topic, which is often referred to as the "municipal bond puzzle."

receive the same tax advantages provided by direct ownership of municipal bonds, and low-cost funds are available with expense ratios as low as 0.1 percent or less.

While mutual funds and exchange-traded funds are compelling options for most investors, certain peculiarities make mutual funds less appealing in the municipal bond market than in other asset classes. With mutual funds, investors are less able to customize their portfolio to their specific state tax considerations. The largest funds often track a national index of municipal bonds. For example, the largest municipal bond fund is currently BlackRock's iShares National Muni Bond ETF, which tracks a national index and has a portfolio that is invested 22 percent in New York and 19 percent in California, with the remainder of the portfolio spread across other US states and the District of Columbia. Similarly, Vanguard's flagship Tax-Exempt Bond Index Fund receives 24 percent of its tax-exempt interest dividends from New York and 15 percent from California. State-specific municipal bond funds make it easier to customize state tax exposure, but they are not prevalent for most states. For example, Vanguard currently offers state-specific funds for six states with high tax benefits—California, Massachusetts, New Jersey, New York, Ohio, and Pennsylvania—but not other states.

A nationally diversified portfolio of municipal bonds is likely suboptimal for almost all investors. Assuming that municipal bond yields in high-tax states are lower because they are priced by in-state investors with large tax benefits (Kidwell, Koch, and Stock 1984; Garrett et al. 2023), it is not clear why investors in other states, who do not receive this state-specific benefit, should hold these bonds. For example, investors in the Vanguard Tax-Exempt Bond Index Fund who reside outside of New York and California miss out on tax benefits equivalent to 19 basis points of yield from these states alone.[5] Compared to the Vanguard fund's expense ratio of as low as 0.09 percent, this represents a significant drag on after-tax performance. One concrete and easily implementable suggestion is that investment management companies could create diversified index funds focused on municipal bonds only from states that do not have state income tax benefits for municipal bonds. While there may be some downsides to a continued shift from direct retail ownership towards greater use of municipal bond funds—during times of financial crisis, Li, O'Hara, and Zhou (2024) show evidence that bond mutual funds decrease the liquidity of this market due to redemptions and forced sales—we think expanded

[5] According to 2023 tax documentation for the Vanguard Tax-Exempt Bond Index Fund, 23.93 percent of tax-exempt interest dividends come from New York, and 15.2 percent of tax-exempt interest dividends come from California. The top marginal tax rate in New York is 14.776 percent (including New York City income taxes), and the top marginal tax rate in California is 14.4 percent. As of August 21, 2024, the fund's 30-day Securities and Exchange Commission yield is 3.34 percent. Thus, the foregone tax benefit from these two states is approximately $0.0334 \times (0.2393 \times 0.14776 + 0.152 \times 0.144) = 0.0019$. Investors in California actually also miss out on this tax benefit because California requires mutual funds to exceed a minimum portfolio allocation percentage to California bonds before being allowed to pass through the California state tax exemption, and this threshold is typically not met by national funds (Vanguard 2024).

use of low-cost funds likely benefits most investors, particularly if the funds are structured to be as tax-efficient as possible.

Could Municipal Bond Tax Treatment Be Changed?

The primary reason for favorable tax treatment of municipal bonds is to subsidize investments by state and local governments. However, the same subsidies could be achieved with direct and calibrated cost-sharing with bond issuers, instead of the somewhat arbitrary subsidy levels that are baked into current tax law; for example, Johnson (2007) and Greenberg (2016) propose changes to municipal bond tax exemptions along these lines. Replacing tax exemptions with direct subsidies to municipalities would eliminate the complicated tax incentives for municipal bond investing and would also likely shift the municipal bond market from being predominantly retail investors to more institutional investors, similar to what we currently see for the corporate bond market, thereby decreasing opportunities for municipal bond dealers to take advantage of retail investors with excessive fees and markups.

Though the politics of removing this tax exemption may be challenging, the US Supreme Court made it clear in *South Carolina v. Baker* (485 U.S. 505 [1988]) that there is no constitutional prohibition to taxing municipal bond interest at the federal level. There is also precedent for taxing municipal bonds; for example, the 2009–2010 Build America Bonds program offered subsidies to municipal bond issuers without tax exemptions for investors (Ang, Bhansali, and Xing 2010b; Cestau, Green, and Schürhoff 2013). Eliminating or limiting the municipal bond tax exemption has been frequently proposed as a way to pay for deficit reduction, spending, or tax cuts going back to the beginning of the exemption in 1918 (as reported by Zweig 2011). But for the same reasons that such proposals have failed in the past, we are skeptical that they will succeed in the future.

Proposals to Improve the Municipal Bond Market

Trading

Overall, requirements for improved post-trade price transparency for municipal bonds have not been effective at reducing transactions costs, as can be seen by the high trading costs for municipal bonds in Figure 2. Moreover, with the large and variable municipal bond markups highlighted in Figures 3 and 4, there appears to be limited enforcement of the "fair pricing" regulations from the Municipal Securities Rulemaking Board. Thus, one policy suggestion would be for the Securities and Exchange Commission to monitor and enforce the MSRB's fair pricing and best execution regulations more strictly, with actions against firms that frequently exhibit large markups. Increased data access and transparency would also be useful. Presently, the MSRB requires a three-year lag before releasing detailed data—and then only with anonymized dealer identifiers. It is hard to understand why the data need to be delayed this long and why identities of the brokers cannot be disclosed. Disclosure of trade-specific markups on a pre-trade basis would be even more

informative, preferably converted to a commission for ease of comparability (Harris 2015). Information along these lines would give customers better visibility into their true trading costs and could put more pressure on dealers to reduce their effective bond commissions, which are typically many multiples of the commissions that the same brokers charge for equities.

Although enhanced enforcement and disclosure could limit examples of more extreme and predatory behavior by brokers, we believe that a dramatic shift in market structure may be necessary to achieve consistently lower transaction costs that are more in line with what investors have come to expect for equities. In this spirit, we echo Harris, Kyle, and Sirri's (2015) discussion and proposal for the Securities and Exchange Commission to require brokers to post customer limit orders to an electronic platform. If customers could access such a platform, they could access the best price quotes from dealers and also trade against one other, as opposed to the current system in which customers can only access quotes from their dealer.

A key component to make such a plan successful is that the Securities and Exchange Commission would need to mandate that brokerage firms display such information to customers (similar to requirements in the 1995 NASDAQ settlement as discussed in Harris, Kyle, and Sirri 2015). At a minimum, information on quotes available from an exchange would inform customers about whether they are receiving the best price on the bonds they are purchasing, and readily-available quote data could also be used by regulators to ensure that brokers are giving their customers best execution, as required by MSRB Rule G-18.

Given the large discrepancies between municipal bond and equity trading costs shown in Figure 2, large decreases in trading costs for municipal bonds seem entirely plausible, as argued by Harris, Kyle, and Sirri (2015).[6] Moving to a centralized marketplace should also foster enhanced liquidity (Biais and Green 2019).

Underwriting

Conflicts of interest and inefficiencies in the underwriting process may always be a challenge in the municipal bond market, due to the highly varied size and sophistication of municipal bond issuers. The recent changes requiring separation between advisors and underwriters are a step in the right direction. State prohibitions on negotiated sales have also been effective at decreasing municipal bond issuance costs (Cestau et al. 2019a), and the market would likely benefit from greater adoption of these policies.

Healthy competition between underwriters is also important to maintain. For example, when Texas banned many national underwriters from participating in the market due to their environmental, social, and governance policies, Garrett and Ivanov (2024) found that costs increased for issuers that had preexisting relationships with the banned underwriters. Lack of competition could be a significant problem more generally if national underwriters continue to withdraw from the

[6] Moreover, access to electronic trading platforms significantly reduced trading costs in the corporate bond market (Hendershott and Madhavan 2015) and in the foreign exchange market (Hau et al. 2021).

market, as some announced in 2023 (Hund et al. 2024). Concentration in the market for underwriters has increased over time, but only for negotiated sales (Cestau 2019), which suggests the possibility that state prohibitions of negotiated sales have an additional benefit of increasing competition (Cestau et al. 2019a).

Standardized disclosure of municipal bond data including fees paid to underwriters, advisors, and credit ratings can also be helpful to let observers learn more about the market and its potential shortcomings. For example, access to underwriter and credit rating fee data from Texas made recent research papers using these data possible and extended our understanding of this market (Luby and Moldogaziev 2013; Cornaggia, Cornaggia, and Israelsen 2023). Conversely, inconsistent accounting standards and limited enforcement make it difficult to assess municipal credit risk (Securities and Exchange Commission 2012, 2020; Carriel 2017). Existing state-level requirements for standardized accounting decrease municipal borrowing costs where they have been implemented (Baber and Gore 2008), but despite a recent Securities and Exchange Commission initiative, muni bonds fail to disclose important information relevant for bond pricing such as the presence of private debt agreements (Ivanov, Zimmermann, and Heinrich 2025). Similarly, state fiscal monitoring policies improve reporting quality and municipal governance more generally (Nakhmurina 2024). If municipalities were required to use financials with uniform accounting principles, it would reduce costs for underwriters, rating agencies, investors, and regulators to assess risk and detect manipulation of financials.

More generally, clearer standardized reporting of key accounting and financial data for municipalities as well as more transparent and standardized reporting of call provisions, derivative reporting, private debt, and fees could decrease the need for having underwriters at all. In other fixed income markets without meaningful risk or asymmetric information—such as Treasury securities and short-term commercial paper—securities are brought to market using auctions and forward contracts without the need for underwriters. In a world in which municipal accounting has greater clarity and consistency, perhaps a similar process would be possible at least for high-quality municipal bonds.

Credit Ratings

Given the complexities of a market with a large share of retail bond investors each trying to evaluate municipal credit risk on their own, there is likely no way around at least some continued reliance on credit ratings, and such ratings do appear to provide useful information to investors (Cornaggia, Cornaggia, and Israelsen 2018). Because the fees paid to credit rating agencies do seem to influence ratings, we agree with the proposal of Cornaggia, Cornaggia, and Israelsen (2023) for the Securities and Exchange Commission to require disclosure of these fees in a manner similar to the required disclosures of auditing fees. Another approach to dealing with biased ratings would be to have at least some ratings move from issuer-paid ratings to investor-paid ratings, which have been shown to be more accurate (Cornaggia and Cornaggia 2013; Bruno, Cornaggia, and Cornaggia 2016).

However, given the disparate ownership of municipal bonds, investors have an incentive to free-ride off ratings once the information is disclosed. A possible proposal is that at bond issuance, issuers could set aside funds that would be directed to an organization acting on behalf of end-investors to pick a rating agency.

A risk of litigation and substantial penalties can serve as a deterrent to nefarious actions. However, as it can take years for the quality of securities to be revealed, and some laws only allow for litigation within three years of the misrepresentation, the statute of limitations may need to be lengthened as was recently done for pandemic relief fraud.

Finally, if financial information for municipal bond issuers used similar accounting principles with standardized reporting in a centralized repository, it would become substantially less costly for independent rating agencies, investor groups, and academics to estimate credit risk models. This would make it easier to analyze individual bonds and the market as a whole, across all issuers and securities. Financial information for bond issuers is available from the Electronic Municipal Market Access (EMMA) website from the Municipal Securities Rulemaking Board (at https://emma.msrb.org), but the information is not currently standardized or machine readable.

Concluding Remarks

A fundamental challenge in the municipal bond market is that disparate issuers and investors often rely on intermediaries with conflicting incentives and superior information. A substantial body of academic research shows the result: hidden fees, unnecessary underwriting costs, high trading costs, and inefficient portfolios that cost issuers and investors alike. Regulatory enhancements by the Municipal Securities Rulemaking Board over the last half-century have generally led to only modest progress, likely indicating that additional regulatory intervention is required.

To make municipal bond trading more competitive, we first propose that the Securities and Exchange Commission could require brokers to give their clients access to external quotes from electronic market platforms similar to the structure of equity markets. Short of this, the SEC and the Municipal Securities Rulemaking Board could much more rigorously enforce current violations of MSRB rules. Second, investors would benefit from more state-specific bond funds as well as more low-cost index funds focused only on states without state income tax advantages. Third, efficient underwriting requires robust, transparent competition. Expanded requirements that municipalities select underwriters through competitive auctions as opposed to negotiated offerings would improve the market. Finally, and perhaps most importantly, the municipal bond market needs more transparency, both in terms of expenses and fees associated with bond issuance and in terms of accounting standards for the municipalities themselves. In a world of more powerful modeling and artificial intelligence tools, data for municipalities with accurate financial information could be more rigorously monitored and modeled with a common

framework across all securities as opposed to the current piecemeal system. In sum, there is hope that the municipal bond market can be brought into the twenty-first century with lower-cost financing for the 50,000 municipal issuers that rely on this $4 trillion market.

■ *We thank Jonathan Parker, Nina Pavcnik, Timothy Taylor, and the rest of the JEP editorial team, as well as Jess Cornaggia, Kimberly Cornaggia, Daniel Garrett, and Andrey Ordin for helpful comments and suggestions. We thank Max Sacher for research assistant support. Griffin is an owner of Integra FEC and Integra Research Group which engage in research, financial consulting, and financial recovery on a variety of issues related to investigating fraud, including various types of bonds. This work was funded by Fundação para a Ciência e a Tecnologia (UIDB/00124/2020, UIDP/00124/2020, UID/00124, Nova School of Business and Economics and Social Sciences DataLab - PINFRA/22209/2016), POR Lisboa and POR Norte (Social Sciences DataLab, PINFRA/22209/2016).*

References

Adelino, Manuel, Sophia Chiyoung Cheong, Jaewon Choi, and Ji Yeol Jimmy Oh. 2023. "Mutual Fund Flows and the Supply of Capital in Municipal Financing." NBER Working Paper 30980.

Adelino, Manuel, Igor Cunha, and Miguel A. Ferreira. 2017. "The Economic Effects of Public Financing: Evidence from Municipal Bond Ratings Recalibration." *Review of Financial Studies* 30 (9): 3223–68.

Adrian, Tobias, Michael Fleming, and Erik Vogt. 2023. "The Evolution of Treasury Market Liquidity: Evidence from 30 Years of Limit Order Book Data." FRB of New York Staff Report 827.

Albright, Amanda. 2024. "Fast-Growing Force in Muni Market Is Upending Mutual Funds' Grip." *Bloomberg*, January 12. https://www.bloomberg.com/news/articles/2024-01-12/fast-growing-force-in-muni-market-is-upending-mutual-funds-grip.

Ang, Andrew, Vineer Bhansali, and Yuhang Xing. 2010a. "Taxes on Tax-Exempt Bonds." *Journal of Finance* 65 (2): 565–601.

Ang, Andrew, Vineer Bhansali, and Yuhang Xing. 2010b. "Build America Bonds." NBER Working Paper 16008.

Ang, Andrew, Richard C. Green, Francis A. Longstaff, and Yuhang Xing. 2017. "Advance Refundings of Municipal Bonds." *Journal of Finance* 72 (4): 1645–82.

Baber, William R., and Angela K. Gore. 2008. "Consequences of GAAP Disclosure Regulation: Evidence from Municipal Debt Issues." *Accounting Review* 83 (3): 565–92.

Babina, Tania, Chotibhak Jotikasthira, Christian Lundblad, and Tarun Ramadorai. 2021. "Heterogeneous Taxes and Limited Risk Sharing: Evidence from Municipal Bonds." *Review of Financial Studies* 34 (1): 509–68.

Baghai, Ramin P., and Bo Becker. 2020. "Reputations and Credit Ratings: Evidence from Commercial Mortgage-Backed Securities." *Journal of Financial Economics* 135 (2): 425–44.

Bagley, John, Stefan Gissler, Kent Hiteshew, and Ivan Ivanov. 2023. "Pushing Bonds over the Edge: Monetary Policy and Municipal Bond Liquidity." http://dx.doi.org/10.2139/ssrn.4330602.

Bagley, John, Marcelo Vieria, and Ted Hamlin. 2022. *Trends in Municipal Securities Ownership*. Municipal Securities Rulemaking Board.

Beatty, Anne, Jacquelyn Gillette, Reining Petacchi, and Joseph Weber. 2019. Do Rating Agencies Benefit from Providing Higher Ratings? Evidence from the Consequences of Municipal Bond Ratings Recalibration." *Journal of Accounting Research* 57 (2): 323–54.

Bergstresser, Daniel. 2023. "The Municipal Bond Market." In *Research Handbook of Financial Markets*, edited by Refet S. Gürkaynak and Jonathan Wright, 301–30. Edward Elgar Publishing.

Bergstresser, Daniel, and Randolph Cohen. 2016. "Changing Patterns in Household Ownership of Municipal Debt." Hutchins Center Working Paper 20.

Bessembinder, Hendrik, Chester Spatt, and Kumar Venkataraman. 2020. "A Survey of the Microstructure of Fixed-Income Markets." *Journal of Financial and Quantitative Analysis* 55 (1): 1–45.

Biais, Bruno, and Richard Green. 2019. "The Microstructure of the Bond Market in the 20th Century." *Review of Economic Dynamics* 33: 250–71.

Bond Dealers of America. 2019. "Time for a Fresh Look at Rule G-23—Benefiting Municipal Issuers, Taxpayers." Bond Buyer Op-Ed, May 16. https://www.bdamerica.org/news-items/bda-presses-for-revision-possible-repeal-of-msrb-rule-g23/.

Brancaccio, Giulia, and Karam Kang. 2022. "Search Frictions and Product Design in the Municipal Bond Market." NBER Working Paper 30775.

Bronshtein, Gila, and Christos A. Makridis. 2020. "The Declining Insurance Benefit in the Municipal Bond Market." *National Tax Journal* 73 (1): 115–56.

Bruno, Valentina, Jess Cornaggia, and Kimberly J. Cornaggia. 2016. "Does Regulatory Certification Affect the Information Content of Credit Ratings?" *Management Science* 62 (6): 1578–97.

Butler, Alexander W. 2008. "Distance Still Matters: Evidence from Municipal Bond Underwriting." *Review of Financial Studies* 21 (2): 763–84.

Butler, Alexander W., Larry Fauver, and Sandra Mortal. 2009. "Corruption, Political Connections, and Municipal Finance." *Review of Financial Studies* 22 (7): 2873–2905.

Cao, Bingkuan (Bryan), Zihan Ye, and Russ Wermers. 2024. "Winning at the Starting Line: Underwriter Connections and Municipal Bond Fund Performance." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4787290.

Carriel, John. 2017. "M-U-N-I: Evidencing the Inadequacies of the Municipal Securities Regulatory Framework." *Business, Entrepreneurship, and Tax Law Review* 1 (2): 472–504.

Cestau, Dario. 2019. "Competition and Market Concentration in the Municipal Bond Market." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.3497599.

Cestau, Dario, Richard C. Green, Burton Hollifield, and Norman Schürhoff. 2019a. "Should State Governments Prohibit the Negotiated Sales of Municipal Bonds?" Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.3508342.

Cestau, Dario, Richard C. Green, and Norman Schürhoff. 2013. "Tax-Subsidized Underpricing: The Market for Build America Bonds." *Journal of Monetary Economics* 60 (5): 593–608.

Cestau, Dario, Burton Hollifield, Dan Li, and Norman Schürhoff. 2019b. "Municipal Bond Markets." *Annual Review of Financial Economics* 11: 65–84.

Chalmers, John, Yu (Steve) Liu, and Z. Jay Wang. 2021. "The Difference a Day Makes: Timely Disclosure and Trading Efficiency in the Muni Market." *Journal of Financial Economics* 139 (1): 313–35.

Chen, Huaizhi, Lauren Cohen, and Weiling Liu. 2024. "Calling All Issuers: The Market for Debt Monitoring." *Management Science*. https://doi.org/10.1287/mnsc.2023.00444.

Cheng, Stephanie F., Christine Cuny, and Hao Xue. 2023. "Disclosure and Competition for Capital." *Management Science* 69 (7): 4312–30.

Christie, William G., and Paul H. Schultz. 1994. "Why Do NASDAQ Market Makers Avoid Odd-Eighth Quotes?" *Journal of Finance* 49 (5): 1813–40.

Christie, William G., and Paul H. Schultz. 1995. "Policy Watch: Did Nasdaq Market Makers Implicitly Collude?" *Journal of Economic Perspectives* 9 (3): 199–208.

Cornaggia, Jess, and Kimberly J. Cornaggia. 2013. "Estimating the Costs of Issuer-Paid Credit Ratings." *Review of Financial Studies* 26 (9): 2229–69.

Cornaggia, Jess N., Kimberly J. Cornaggia, and John E. Hund. 2017. "Credit Ratings across Asset Classes: A Long-Term Perspective." *Review of Finance* 21 (2): 465–509.

Cornaggia, Jess, Kimberly J. Cornaggia, and Ryan Israelsen. 2023. "Rating Agency Fees: Pay to Play in Public Finance?" *Review of Financial Studies* 36 (5): 2004–45.

Cornaggia, Jess, Kimberly J. Cornaggia, and Ryan D. Israelsen. 2018. "Credit Ratings and the Cost of Municipal Financing." *Review of Financial Studies* 31 (6): 2038–79.

Cornaggia, Jess N., Kimberly J. Cornaggia, and Ryan D. Israelsen. 2020. "Where the Heart Is: Information

Production and the Home Bias." *Management Science* 66 (12): 5532–57.

Cornaggia, Kimberly, John Hund, and Giang Nguyen. 2022. "Investor Attention and Municipal Bond Returns." *Journal of Financial Markets* 60: 100738.

Cornaggia, Kimberly, John Hund, and Giang Nguyen. 2024. "The Price of Safety: The Evolution of Municipal Bond Insurance Value." *Management Science* 70 (4): 2330–54.

Faulk, Kent. 2015. "Prison Sentence Ends for JeffCo Sewer Contractor; 3 More to Go." *Advance Local*, October 30. https://www.al.com/news/birmingham/2015/10/jeffco_sewer_contractor_ends_p.html.

Federal Reserve. 2025. *Financial Accounts of the United States—Z.1: L.212 Municipal Securities, Series FL893062005.* https://www.federalreserve.gov/releases/z1/preview/html/l212.htm (accessed July 17, 2024).

Garrett, Daniel G. 2024. "Conflicts of Interest in Municipal Bond Advising and Underwriting." *Review of Financial Studies* 37 (12): 3835–76.

Garrett, Daniel G., and Ivan T. Ivanov. 2024. "Gas, Guns, and Governments: Financial Costs of Anti-ESG Policies." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4123366.

Garrett, Daniel, Andrey Ordin, James W. Roberts, and Juan Carlos Suárez Serrato. 2023. "Tax Advantages and Imperfect Competition in Auctions for Municipal Bonds." *Review of Economic Studies* 90 (2): 815–51.

Gillette, Jacquelyn R., Delphine Samuels, and Frank S. Zhou. 2020. "The Effect of Credit Ratings on Disclosure: Evidence from the Recalibration of Moody's Municipal Ratings." *Journal of Accounting Research* 58 (3): 693–739.

Green, Richard C. 2007. "Presidential Address: Issuers, Underwriter Syndicates, and Aftermarket Transparency." *Journal of Finance* 62 (4): 1529–50.

Green, Richard C., Burton Hollifield, and Norman Schürhoff. 2007a. "Dealer Intermediation and Price Behavior in the Aftermarket for New Bond Issues." *Journal of Financial Economics* 86 (3): 643–82.

Green, Richard C., Burton Hollifield, and Norman Schürhoff. 2007b. "Financial Intermediation and the Costs of Trading in an Opaque Market." *Review of Financial Studies* 20 (2): 275–314.

Green, Richard C., Dan Li, and Norman Schürhoff. 2010. "Price Discovery in Illiquid Markets: Do Financial Asset Prices Rise Faster Than They Fall?" *Journal of Finance* 65 (5): 1669–1702.

Greenberg, Scott. 2016. "Reexamining the Tax Exemption of Municipal Bond Interest." Tax Foundation Fiscal Fact 520.

Griffin, John M. 2021. "Ten Years of Evidence: Was Fraud a Force in the Financial Crisis?" *Journal of Economic Literature* 59 (4): 1293–1321.

Griffin, John M., Nicholas Hirschey, and Samuel Kruger. 2023. "Do Municipal Bond Dealers Give Their Customers 'Fair and Reasonable' Pricing?" *Journal of Finance* 78 (2): 887–934.

Griffin, John M., Nicholas Hirschey, and Samuel Kruger. 2025. *Data and Code for: "Why Is the Fragmented Municipal Bond Market So Costly to Investors and Issuers?"* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E222541V1.

Griffin, John M., and Jordan Nickerson. 2023. "Are CLO Collateral and Tranche Ratings Disconnected?" *Review of Financial Studies* 36 (6): 2319–60.

Griffin, John M., and Dragon Yongjun Tang. 2012. "Did Subjectivity Play a Role in CDO Credit Ratings?" *Journal of Finance* 67 (4): 1293–1328.

Harris, Larry. 2015. "Transaction Costs, Trade Throughs, and Riskless Principal Trading in Corporate Bond Markets." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.2661801.

Harris, Larry, Albert S. Kyle, and Erik R. Sirri. 2015. "Statement of the Financial Economists Roundtable, April 2015: The Structure of Trading in Bond Markets." *Financial Analysts Journal* 71 (6): 5–8.

Harris, Lawrence E., and Michael S. Piwowar. 2006. "Secondary Trading Costs in the Municipal Bond Market." *Journal of Finance* 61 (3): 1361–97.

Hau, Harald, Peter Hoffmann, Sam Langfield, and Yannick Timmer. 2021. "Discriminatory Pricing of Over-the-Counter Derivatives." *Management Science* 67 (11): 6660–77.

Hendershott, Terrence, and Ananth Madhavan. 2015. "Click or Call? Auction versus Search in the Over-the-Counter Market." *Journal of Finance* 70 (1): 419–47.

Herpfer, Christoph, and Gonzalo Maturana. 2021. "Who Prices Credit Rating Inflation?" Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.3579030.

Howard, Cooper. 2024. "When to Consider Munis from Outside Your Home State." Charles Schwab, October 30. https://www.schwab.com/learn/story/when-to-choose-munis-from-outside-your-home-state.

Hund, John, Christian Lundblad, Christos Makridis, and Giang Nguyen. 2024. "The Rise of Investor Sophistication and the Decline of Underwriting Profits in the Municipal Bond Market." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.4867270.

IRS. 2025. "Tax-Exempt Bonds." Statistics of Income Division. https://www.irs.gov/statistics/soi-tax-stats-tax-exempt-bond-statistics.

Ivanov, Ivan T., Tom Zimmermann, and Nathan W. Heinrich. 2025. "Limits of Disclosure Regulation in the Municipal Bond Market." *Management Science*. https://doi.org/10.1287/mnsc.2022.02289.

Johnson, Calvin H. 2007. "Repeal Tax Exemption for Municipal Bonds." *Tax Notes* 117: 1259.

Jones, Charles M. 2002. "A Century of Stock Market Liquidity and Trading Costs." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.313681.

Kidwell, David S., Timothy W. Koch, and Duane R. Stock. 1984. "The Impact of State Income Taxes on Municipal Borrowing Costs." *National Tax Journal* 37 (4): 551–61.

Landoni, Mattia. 2018. "Tax Distortions and Bond Issue Pricing." *Journal of Financial Economics* 129 (2): 382–93.

Li, Dan. 2007. "New Findings in the Market Microstructure of Over-the-Counter Markets." PhD diss., Carnegie Mellon University.

Li, Dan, and Norman Schürhoff. 2019. "Dealer Networks." *Journal of Finance* 74 (1): 91–144.

Li, Yi, Maureen O'Hara, and Xing (Alex) Zhou. 2024. "Mutual Fund Fragility, Dealer Liquidity Provision, and the Pricing of Municipal Bonds." *Management Science* 70 (7): 4802–23.

Luby, Martin, and W. Bartley Hildreth. 2014. "A Descriptive Analysis of the Municipal Advisors Market." *Municipal Finance Journal* 34 (4): 69–98.

Luby, Martin, and Tima Moldogaziev. 2013. "An Empirical Examination of the Determinants of Municipal Bond Underwriting Fees." *Municipal Finance Journal* 34 (2): 13–50.

Malakar, Baridhi. 2024. "Fiduciary Duty in the Municipal Bonds Market." Preprint, arXiv. https://doi.org/10.48550/arXiv.2406.15197.

Municipal Securities Rulemaking Board (MSRB). 2016. "New Disclosure Requirements under MSRB Rule G-15 and Prevailing Market Price Guidance Pursuant to Rule G-30 Effective May 14, 2018." Regulatory Notice 2016-28, November 29. https://www.msrb.org/sites/default/files/2016-28.pdf.

Nakhmurina, Anya. 2024. "Does Fiscal Monitoring Make Better Governments? Evidence from U.S. Municipalities." *Accounting Review* 99 (4): 395–425.

Raineri, Lori, Mark Robbins, Bill Simonsen, and Keith Weaver. 2012. "Underwriting, Brokerage, and Risk in Municipal Bond Sales." *Municipal Finance Journal* 33 (2): 87–103.

Schleicher, David. 2023. *In a Bad State: Responding to State and Local Budget Crises*. Oxford University Press.

Schultz, Paul. 2012. "The Market for New Issues of Municipal Bonds: The Roles of Transparency and Limited Access to Retail Investors." *Journal of Financial Economics* 106 (3): 492–512.

Securities and Exchange Commission. 2012. *Report on the Municipal Securities Market*. US Securities and Exchange Commission.

Securities and Exchange Commission. 2015. "In the Matter of Edward D. Jones & Co., L.P." Securities Act of 1933 Release 9889, August 13; Securities Exchange Act of 1934 Release 75688, August 13; Administrative Proceeding File 3–16751. https://www.sec.gov/files/litigation/admin/2015/33-9889.pdf.

Securities and Exchange Commission. 2016. "Press Release: SEC Completes Muni-Underwriter Enforcement Sweep: 72 Firms Charged since June 2015." February 2. https://www.sec.gov/newsroom/press-releases/2016-18.

Securities and Exchange Commission. 2020. *Preliminary Recommendation Regarding Timeliness of Financial Disclosures in the Municipal Securities Market*. US Securities and Exchange Commission.

Selway, William. 2013. "JPMorgan's Alabama Debacle Set to Cost Bank $1.6 Billion." *Bloomberg*, June 5. https://www.bloomberg.com/news/articles/2013-06-05/jpmorgan-s-alabama-debacle-set-to-cost-bank-1-5-billion.

Vanguard. 2024. "Tax-Exempt Interest Dividends by State for Vanguard Municipal Bond Funds and Vanguard Tax-Managed Balanced Fund." https://investor.vanguard.com/content/dam/retail/publicsite/en/documents/taxes/inbst-2024.pdf.

Zweig, Jason. 2011. "How Long Willl the Tax Break on Municipal Bonds Last" *Wall Street Journal*, May 7. https://www.wsj.com/articles/SB10001424052748704810504576307233579693982.

Journal of Economic Perspectives—Volume 39, Number 1—Winter 2025—Pages 261–274

Retrospectives: Yair Mundlak and the Fixed Effects Estimator

Marc F. Bellemare and Daniel L. Millimet

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).

Unobserved heterogeneity is the bane of the applied econometrician's existence. It arises from omitted variables, or the existence of factors which the econometrician does not observe and cannot control for, but which are correlated with variables that the econometrician does observe and control for, leading to biased estimates. In labor economics, for example, individuals choose whether to go to college in part based on their unobservable expectations about how a college degree will increase their future earnings. But if those more likely to obtain a college degree expect greater earnings, then any estimate of the relationship between one's earnings and one having a college degree will be biased. Similarly, in development economics, farmers often choose whether to adopt a new technology that can increase yields in part based on their unobservable risk preferences. But if those more likely to adopt a new technology are less likely to be

■ *Marc F. Bellemare is McKnight Presidential Chair in Applied Economics, Distinguished McKnight University Professor, Distinguished University Teaching Professor, and Northrop Professor, University of Minnesota, St. Paul, Minnesota. Daniel L. Millimet is Robert H. and Nancy Dedman Trustee Professor, Southern Methodist University, Dallas, Texas, and Research Fellow, IZA—Institute of Labor Economics, Bonn, Germany. Their email addresses are mbellema@umn.edu and millimet@smu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20241406.

risk-averse, then any estimate of the relationship between one's yields and whether one has adopted a new technology will also be biased.

In a field or laboratory experiment, the problem of unobservable variables is sidestepped by random assignment. That is, the researcher assumes that unobserved variables are distributed the same way between treatment and control groups. But when a randomized experiment is unavailable and an applied economist must make do with observational data, the danger of biased results from unobserved variables is always lurking.

In the toolkit of the applied economist, the fixed effects estimator is one of the methods available to deal with unobserved variables in observational data. This approach was first applied in the literature by Israeli agricultural economist Yair Mundlak (1961) in a seminal article published in the *Journal of Farm Economics* (which was renamed *American Journal of Agricultural Economics* in 1968).

We begin with a brief biography of Mundlak. We then take a step back to look at the historical statistical and economic contexts in which Mundlak applied the fixed effects estimator. We then discuss Mundlak's original application that brought the fixed effects estimator into economics: the study of agricultural productivity, with the use of farm fixed effects to control for farm management quality. We briefly discuss how, in the choice between fixed and random effects, the fixed effects estimator won the day and became the estimator of choice among applied economists. In the conclusion, we point out that the needed elements for the fixed effects estimator were already there in the work of earlier statisticians. In addition, the fixed effects estimator has experienced a renaissance over the past few decades, starting with the "credibility revolution" (in this journal, see Angrist and Pischke 2010) and continuing with the increased popularity of difference-in-differences designs and two-way fixed effects estimators.

Biography

Yair Mundlak was born in 1927 in Pinsk, a city in the Brest region which was then part of Poland, but now part of Belarus. In the years prior to World War II, the Mundlak family emigrated to the area that is now part of Israel, but at the time was part of the British Mandate, where a young Yair attended Kadoorie Agricultural High School in Lower Galilee for his last two years of high school.

After graduating from high school, Mundlak fought in the 1948 Arab-Israeli War, after which he joined a kibbutz, or collective farm, which he ended up managing. It was then that Mundlak realized that he might benefit from a university education (Mundlak 2011). Students from Kadoorie were not eligible to study at the Hebrew University of Jerusalem—the obvious choice at the time in Israel for someone interested in studying agriculture—and so in 1950, Mundlak left for the United States to study at what was then the College of Agriculture at Davis, an extension school for the University of California at Berkeley, which later became the University of California-Davis.

After graduating in 1953 with a BSc (Highest Honors) in Agricultural Economics, Mundlak headed to the University of California at Berkeley for his graduate studies, first earning an MS in Statistics in 1956 and then a PhD in Agricultural Economics in 1957. He won the outstanding PhD dissertation award from the American Farm Economics Association (now the Agricultural and Applied Economics Association).

Mundlak then returned to Israel to join the faculty of the Hebrew University of Jerusalem as associate professor and, starting in 1970, as full professor. He chaired the Department of Agricultural Economics at Hebrew University from 1965 to 1970, founding the Center for Agricultural Economic Research in 1968. From 1972 to 1974, he served as Dean of the School of Agriculture at Hebrew University. In 1978, he joined the Department of Economics at the University of Chicago, where he held the F. H. Prinze Chair in Economics until 1997. He died on October 20, 2015, at the age of 88, survived by his wife Yaffa and his children Guy, Tal, and Yael.

Farm Productivity and the Confounding Problem of Management Quality

During the 1940s and 1950s, a number of economists were interested in explaining differences in agricultural productivity. Management had been acknowledged as an omitted variable in the estimation of farm production functions as far back as work by Tintner (1944) and Heady (1946). While this focus on farms may strike modern-day readers as quaint or outdated, a number of seminal econometric contributions originated in looking at applications in agricultural and resource economics—in part because agricultural statistics were among the first collected historically by governments (Scott 1999).[1] Much of this section relies on the excellent article by Dupont-Kieffer and Pirotte (2011), who recount the history of the early years of panel-data econometrics, and to whom we owe a considerable debt of gratitude.

Marschak and Andrews (1944) were interested in the estimation of firm-specific production functions. In doing so, they wanted to isolate the differences between firms due to the use of different technologies as well as the differences between firms due to time. They pointed out the fundamental identification problem involved in estimating production functions: specifically, inputs are not assigned at random, which means that estimates of elasticities of output with respect to inputs can be biased, as are estimates of returns to scale. They also discussed the notion of "interfirm" production functions ("functions fitted to the data on the output,

[1] Prominent examples, among others, would include Sir Ronald Fisher and his (literal) field experiments, Philip Wright with instrumental variables, Frederick Waugh with the Frisch-Waugh-Lovell theorem, and Marc Nerlove with random effects. See Fox (1986) and Bessler et al. (2010) for discussions of the contributions of agricultural economists to econometrics.

manpower, and capital of a number of individual firms"), noting expressed doubts about whether such "interfirm" production functions could be identified.

Hoch (1957, 1958, 1962) studied the same problem as Marschak and Andrews (1944), but he combined cross-sectional and time-series data in an effort to control for unobserved heterogeneity, and he concluded that "management" was an important omitted variable. The earliest attempt at empirically controlling for management of which we are aware was made by Reiss (1952) who, in his PhD dissertation at the University of Illinois, estimated a Cobb-Douglas production function for a cross-section of farms "controlling" for management by including a coarse proxy for whether a given farm operator was a "good" or "poor" farmer. Given this coarse proxy, or management measured with error, it is perhaps unsurprising that the management input did not play a statistically significant role in explaining productivity in his study.[2] Hoch (1957) also attempted to control for management using fixed effects in his PhD dissertation at the University of Chicago, although his paper was not published until after Mundlak's (1961) work. [3]

When Griliches (1957) wrote his famous paper on "specification bias," he showed how, under the relatively uncontroversial assumption that (unobserved) management is positively correlated with the (observed) inputs, omitting management would bias the estimates of the various elasticities of output with respect to inputs upward and the estimates of returns to scale downward. Griliches (p. 8) added: "We underestimate returns to scale if we exclude an input that varies less than proportionately with the included inputs, and vice versa . . . under 'reasonable' assumptions the omission of managerial inputs from the production function biases the estimate of the elasticity of output with respect to capital inputs upwards and the estimate of returns to scale downwards." He further (p. 13) emphasized the importance of managerial quality as a specification error: "The specification error conceded most often by estimators of production functions is the omission of . . . managerial services . . . I am aware of only one published attempt to include

[2] For a discussion of how Reiss developed his proxy for management, see Reiss (1949). For a recent survey of the literature looking at the effect of management on productivity, see Papadopoulos (2022).

[3] While Irving Hoch is sometimes credited as the "father of fixed-effect modeling" (see the memorial at https://epps.utdallas.edu/irv-hoch/), his paper on the topic (Hoch 1962) was not published until after Mundlak published his 1961 article. This may explain why, in the minds of most economists, Mundlak (1961) is associated with the fixed effects estimator. Hoch did publish a short half-page progress report on his dissertation in *Econometrica* in 1955 (Econometric Society 1955, pp. 325–26) as part of a report on the previous meetings of the Econometric Society, which had been held in Montreal in 1954. In that progress report, Hoch reports some estimation results for a Cobb-Douglas production function on 63 Minnesota farms observed over six years. In those results, Hoch includes both a two-way (farm and year) fixed effects specification, as well as a specification with year fixed effects. For the purposes of this article, we will follow the literature and treat Hoch's 1955 progress report as not being part of the published literature when Mundlak published his own 1961 article (indeed, even to this day Hoch's earlier report is not part of the canon of panel-data econometrics), even though Mundlak cites Hoch's progress report in his landmark article. That Mundlak appears to have beaten Hoch to the punch as far as publishing a full original research article does not seem to have caused any animosity between the two, since they subsequently coauthored an article on the consequences of alternative specifications of a Cobb-Douglas production function (Mundlak and Hoch 1965).

managerial services in the production function . . . In a footnote to a recent [chapter], Earl R. Swanson reports on an attempt by F.G. (sic) Reiss at the University of Illinois to add an index of 'managerial ability' or 'quality' to the production function."

A Concrete Application of Fixed Effects

Mundlak was interested in obtaining unbiased estimates of the effects of observed inputs on agricultural output. He did recognize, however, that the effect of management could be a confounding factor. For example, if generally well-managed firms were more likely to increase the use of certain inputs or to use other productivity-improving practices, then a regression estimate of the correlation between these inputs and output would be biased upward, because it would mix the effects of the input and of good management. Mundlak (1961, p. 44) wrote: "It has been felt for a long time that the estimates of the parameters of production functions are subject to bias as a result of excluding the variable which represents management." To address the problem, Mundlak (p. 44) assumed that "whatever management is, it does not change considerably over time; and for short periods, say a few years, it can be assumed to remain constant." Thus, with data on at least two time periods within a short interval, it should be possible to hold management constant, thereby purging the error term of it and its correlation with the inputs on the right-hand side of his regression equation. Specifically, Mundlak begins with the following regression equation:

$$Y_{it} = B_0 + B_1 X_{1it} + \ldots + B_k X_{kit} + CM_i + e_{it},$$

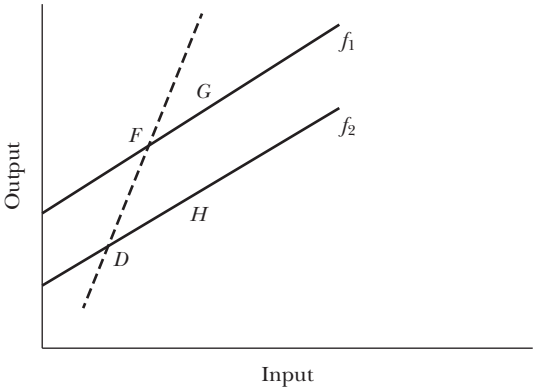
where Y denotes output, the X variables denote observed inputs, M denotes management, and e is a mean-zero error term, with $i = 1, \ldots, N$ and $t = 1, \ldots, T$. In such a setup, $(B_1, \ldots, B_k, C)$ is the vector of population parameters to be estimated.[4]

Mundlak (1961) then points out that, because the management input variable M is unobserved and no adequate proxy is available, business as usual consists of using cross-sectional data and estimating a version of this equation that excludes M and omits the t subscripts, given the cross-sectional nature of the data, all of which will bias the estimates of the coefficients, because the management variable will co-vary with the use of other inputs X.

We reproduce Mundlak's (1961) visual demonstration of the problem in Figure 1. Here, f_1 and f_2 show output as a function of a single input for firms 1 and 2, respectively. Both production functions have the same slope, measured either by the slope between points F and G or between points D and H, but their intercepts differ because of management, with firm 1 having better management than firm 2.

[4] While modern-day readers will likely be used to lower-case Latin letters used to denote variable names, and while they will certainly be more used to seeing Greek letters used to denote parameters, we use the notation in Mundlak (1961).

Figure 1
Bias Arising from a Failure to Account for Management



Source: Mundlak (1961).
Note: The lines for f_1 and f_2 represent production functions for two different firms. The relationship between changes in measured input and resulting changes in output is the same between the firms, as shown by the same slope. Firm 1, however, has better management than firm 2. The dashed line shows the result of estimating data from both firms, which clearly differs from the slope of either f_1 or f_2.

Ignoring management and estimating a single production function for both firms yields a biased estimate of the slope that is equal to the slope between points *F* and *D*, and that clearly differs from the (true) slopes of either f_1 or f_2.

Mundlak explained that the key to estimating the slope of f_1 and f_2 was to have two points on each of those production functions—something which could be done by collecting data for two consecutive years and then combining cross-sectional with time-series data with the assumption that the production function for each firm was constant across the two time periods. Thus, using annual data on 66 Israeli farms for the period 1954–1958, he regresses the value of each farm's output in a year on the number of labor days (X_1), variable expenses (X_2), the value of livestock at the beginning of the year (X_3), the value of livestock and poultry barns (X_4), the amount of irrigated land (X_5), and the fixed effect CM_i, which he rewrites as A_i since only the product of C and M_i is identified in the absence of observed data on M_i.[5] For estimation, Mundlak took the logarithm of each variable in X, subtracted the within-farm mean of each variable, and estimated B_0, B_1, . . ., B_k, and A, effectively using variation within each farm over time (which is why fixed effects is sometime called "the within estimator") to estimate a Cobb-Douglas production function.[6]

[5] The fixed effects estimator is frequently referred to as the "within" estimator. It is also referred to less frequently as the least squares dummy variables estimator or the mean-differencing estimator.
[6] Recall that if output Q is produced using capital K and labor L such that $Q = F(K, L)$, imposing a Cobb-Douglas functional form means that $Q = AK^{\alpha} L^{\beta}$, where A is total factor productivity and α and β are elasticities whose sum $\alpha + \beta$ is a measure of returns to scale. Taking logarithms on both sides of the

Table 1
Mundlak's (1961) Empirical Results

Estimated elasticities	*Pooled ordinary least squares* (1)	*With year fixed effects* (2)	*With farm fixed effects* (3)	*With two-way fixed effects* (4)
X_1 (number of labor days)	0.130**	0.153**	0.083*	0.114**
X_2 (variable costs)	0.692**	0.679**	0.635**	0.582**
X_3 (value of livestock)	0.004	0.004	0.002	0.005
X_4 (value of livestock and poultry barns)	0.103**	0.101**	0.156**	0.100*
X_5 (amount of irrigated land)	0.037**	0.032*	0.002	−0.007

Note: * and **, respectively, denote significance at the 5 and 1 percent level.

Table 1 reproduces Mundlak's empirical results, with column 1 showing results for a business-as-usual cross-sectional (that is, pooled ordinary least squares) approach omitting farm fixed effects, column 2 omitting farm fixed effects but including year fixed effects, column 3 including farm fixed effects but omitting year fixed effects, and column 4 including both farm and year fixed effects. This latter specification aimed to control for both management as well as for the changing nature of what in the production function is shared by all the farms in his sample in each time period he considers. For example, while farm fixed effects could control for differences in management, year fixed effects could account for weather to the extent that the farms in Mundlak's sample were all similarly affected by the same weather. To our knowledge, by implicitly recognizing that fixed effects could be used to control for more than one source of unobserved heterogeneity, Mundlak was first to publish the results of a two-way fixed effects estimator, which is now widely used in applied work. This additional contribution of Mundlak's seems to have been overshadowed by his core contribution.

In particular, Mundlak's (1961) results show that accounting for farm fixed effects—which he viewed solely as capturing the effect of management (rather than, say, physical differences between agricultural land like superior natural irrigation or drainage that also would not vary between years)—led to results that differed significantly from those obtained from pooling various cross-sections. Comparing the results in column 1 (pooled ordinary least squares) and 3 (the farm fixed effects) of Table 1, nearly all of the estimated elasticities decrease in magnitude when including farm fixed effects (the only exception is the elasticity of output with respect to the value of livestock and poultry barns, X_4). Intuitively, failure to account for management as an input in column 1—by omitting farm fixed effects—means that the effect of management is picked up by those inputs that are included in column 1, and so the results in column 1 overestimate elasticities compared with those in column 3. Likewise, the inclusion of year fixed effects in addition to farm fixed effects in column 4

equation yields $\ln Q = \ln A + \alpha \ln K + \beta \ln L$. In its empirical version, the first term of this last equation becomes the intercept, or constant, and an error term is added.

further decreases the magnitude of four out of five of the estimated elasticities, going so far as to flip the sign of the estimated elasticity of output with respect to irrigated land while leaving that same elasticity statistically insignificant.

Again, Mundlak (1961) was not the first economist to look at a fixed effects estimator. Hildreth (1950) had already done so in a Cowles Commission paper titled "Combining Cross Section Data and Time Series." What Mundlak did was to show other economists how useful the fixed effects estimator could be, by showing how much of a difference it made relative to a pooled ordinary least squares estimator, as well as by showing how specific economic attributes could be ascribed to the unobserved heterogeneity captured by the fixed effect. In other words, it was finding a good application that tied Mundlak's name to the fixed effects estimator.

Fixed Effects and Random Effects Estimators

The early years of panel-data econometrics in the 1960s were characterized by two seminal articles that did much to set the tone of the conversation. The first article is Mundlak (1961), which brought the fixed effects estimator to economics. The second article is Balestra and Nerlove (1966), which brought the random effects estimator to economics. Thus, Dupont-Kieffer and Pirotte (2011) write of Pietro Balestra (1935–2005), Yair Mundlak (1927–2015), and Marc Nerlove (1933–) as the founding parents of panel-data econometrics.[7] In much of the empirical literature on panel data over the following decades, one of the first questions faced by applied econometricians was whether to use fixed or random effects.

What is the random effects estimator and how does it differ from fixed effects? Importantly, the original distinction between fixed and random effects was conceptual and not directly related to estimation. Wooldridge (2010, p. 285) describes the difference: "[O]ne often sees a discussion about whether [the unit-specific effect] will be treated as a random effect or a fixed effect. Originally, such discussions centered on whether [the unit-specific effect] is properly viewed as a random variable or as a parameter to be estimated. In the traditional approach to panel-data models, [the unit-specific effect] is called a 'random effect' when it is treated as a random variable and a 'fixed effect' when it is treated as a parameter to be estimated for each cross section observation *i*."

The first published application of the random effects estimator was from Balestra and Nerlove (1966), looking at the market for natural gas. They arrived at the use of random rather than fixed effects because of the inclusion of lagged output (that is, Y_{it-1}) as a regressor, and because the fixed effects estimator led to

[7] For readers interested in a broader discussion of the early years of panel-data econometrics, we strongly recommend Dupont-Kieffer and Pirotte (2011), along with Nerlove's (2005) chapter titled "The History of Panel Data Econometrics" in his collection of essays on econometrics. We also owe a debt of another nature to Dupont-Kieffer and Pirotte (2011): Because some of the references we cite in this article were not accessible to us, we rely on those authors' discussion of those sources when citing them. To alleviate the text and not burden the reader, however, we do not flag instances of such vicarious citations.

results they deemed "highly implausible" (p. 592). They wrote: "The presence of lagged endogenous variables may make it difficult, if not impossible, to separate the individual (state) effects from the effect induced by the lagged variable."[8]

A lot of early discussion concerning whether an econometrician should use fixed or random effects with panel data had to do with whether one had access to a whole population (say, all 50 US states over time), in which case treating the unobserved effects as fixed parameters to be estimated was thought preferable, rather than a random sample of the population (say, a random sample of the US population over time), in which case treating the unobserved effects as random draws a distribution to be estimated was thought preferable.[9]

In a 1978 *Econometrica* article, Mundlak argued that focusing on the nature of the effect—fixed or random—was a reasonable starting point but also an inadequate approach. Instead, he argued that the more important issue is whether or not the effects are correlated with the included variables in the regression model. Mundlak then unified the two approaches through what has now become known as the "Mundlak approach." To do so, he started by decomposing the error term e_{it} in the earlier equation into three components: (1) a systematic unit i-specific error term, denoted m_i; (2) a systematic time period t-specific error term, denoted s_t; and (3) an error term free of unit- and time-specific effects and unique to each observation u_{it}, such that $e_{it} = m_i + s_t + u_{it}$. He then explicitly modeled the relationship between the unit-specific error term m and the time specific error term s and the included variables, X. The model also included a random component. Mundlak showed that the fixed and random effects estimators are both generalized least squares estimators, where the latter is a restricted version of the former as it omits any relationship between m (and s) and X in the auxiliary model of the effects.

On this issue, Wooldridge (2010, p. 286) writes: "[T]he key issue involving [the unit-specific effect] is whether or not it is uncorrelated with the observed explanatory variables . . . Mundlak (1978) made this argument many years ago, and it still is persuasive." This practical difference was reinforced with the publication of Hausman (1978) in the same year, after which the Durbin-Wu-Hausman specification test, which pits the two estimators against one another, became the standard for discriminating between the two.

The random effects estimator does make an unpalatable assumption, however; that is, it assumes strict exogeneity of the error term (inclusive of the unobserved heterogeneity) with respect to the right-hand side variables—an assumption which is at best heroic and at worst downright untenable in the absence of experimental data. Even with experimental data, the random effect estimator imposes more structure on the error term than many practitioners deem necessary. The end result is that the fixed effects estimator largely has won the day, at least in the constituent

[8] Nickell (1981) subsequently showed how ordinary least squares estimates are biased when including both fixed effects and an autoregressive term.

[9] Indeed, one of us was taught this very heuristic at the Université de Montréal as recently as the late 1990s.

fields of applied microeconomics—say, agricultural and resource, development, environmental, health, and labor economics—so much so that specification tests of the type proposed in Hausman (1978) are often considered unnecessary.

Aftermath

We have shown how, contrary to a common misconception, Mundlak (1961) did not derive the fixed effects estimator nor, for that matter, did Balestra and Nerlove (1966) derive the random effects estimator.[10] By the time those two articles were published, those estimators were available "off the shelf" from the extant statistical literature. As Dupont-Kieffer and Pirotte (2011) point out, Fisher (1925) introduced the concept of fixed effects, Airy (1861) and Chauvenet (1871) had already introduced the concept of random effects, and Daniels (1939) clarified the difference between the two. While Mundlak clearly made important contributions, it was finding a compelling application that led to Mundlak's name being associated with the fixed effects estimator among economists.

In recent years, more than 60 years since the publication of Mundlak (1961), the fixed effects estimator seems more relevant than ever. Hill et al. (2020, p. 367) state that "fixed-effects models for panel data are now widely recognized as powerful analytic tools for longitudinal data analysis," and Imai and Kim (2021, p. 405) write that "linear regression with unit and time fixed effects [is] the default methodology for estimating causal effects from panel data."

In particular, the fixed effects estimator is central to two popular modern methodologies. In a difference-in-differences design, units are followed over time, some of which get treated and some of which do not, and the variation over time and across units is used to get at the causal effect of treatment. In two-way fixed effects, unit and time fixed effects are used to account for time-invariant heterogeneity within each unit and unit-invariant heterogeneity within each time period, and thus to get at the causal effect of treatment. Both approaches require specific assumptions, which naturally and appropriately come under close scrutiny. Not a season goes by without its bountiful harvest of new working papers calling into question the estimates obtained from difference-in-differences or two-way fixed effects. Useful starting points for entering this literature include Callaway, Goodman-Bacon, and Sant'Anna (2024), de Chaisemartin and d'Haultfœuille (2020, 2023), Imai and Kim (2021), Jakiela (2021), Miller (2023), Sun and Shapiro (2022), and Wooldridge (2021). A common concern is the recognition that the fixed effects estimator can produce misleading estimates when heterogeneity exists in a model's population parameters.

[10] Even some of those who knew Mundlak well tend to assume that he derived the fixed effects estimator. In his remembrance of Mundlak, Zilberman (2016) wrote that "[Mundlak] developed a method called 'fixed effect estimation procedure' to identify how yield varies among different farmers, villages and over different seasons."

Another issue is that as researchers gain access to longer and longer panel-data sets, there is less and less heterogeneity remaining constant over time. While this point is sometimes lost in current applied research, it was not lost on Mundlak (1978), who wrote: "[I]t would be unrealistic to assume that the individuals do not change in a differential way as the model assumes . . . [I]t is more realistic to assume that individuals do change differentially but at a pace that can be ignored for short time intervals." In Millimet and Bellemare (2024), we present a number of alternative estimators that fare better than the fixed effect estimator in the context of simulations and replications of published articles.

But with such concerns duly noted, new applications for fixed estimators are arising as well. Rather than merely being a means to controlling for unobservable variables, the fixed effects are now often themselves the object of interest. Well-known examples include the analysis of worker and firm fixed effects in matched employee-employer data that originated in the work of Abowd, Kramarz, and Margolis (1999), which addresses questions like whether certain firms pay premium wages for similar experience. Another example is the analysis of estimating the magnitude of teacher quality using matched teacher-student data and fixed effects for teachers, as in Rivkin, Hanushek, and Kain (2005). Yet another example is the use of judge fixed effects, in which the effects of judicial decisions (like length of incarceration sentence granting bail) on offenders can be evaluated because otherwise similar defendants are, in effect, randomly assigned to judges with systematically different decisions, as in Kling (2006). Another recent development—now beginning to take hold in applied research—aggregates observations into groups and includes group fixed effects in the model. Such approaches can also allow for the possibility that group fixed effects and even the group structure may change after a few periods (for example, Bonhomme and Manresa 2015; Bonhomme, Lamadon, and Manresa 2022; Lumsdaine, Okui, and Wang 2023).[11] Given that, the fixed effects estimator seems likely to remain part and parcel of the applied econometrician's toolkit.

■ *We thank the JEP editors for encouraging us to write this article, as well as Jonathan Parker, Nina Pavcnik, and Timothy Taylor for their comments on an earlier version. We are also grateful to Danielle Becker-Rosenbaum for her help with locating important bibliographical resources. Bellemare would like to dedicate this article to the memory of Claude Montmarquette (1942–2021), who awakened his interest in applied econometrics.*

[11] Moreover, the Mundlak approach is now recognized as a way of dealing with the incidental parameter problem that often arises in nonlinear models with fixed effects. Indeed, there is now a "mundlak" command in Stata. See http://fmwww.bc.edu/RePEc/bocode/ m/mundlak.html.

References

Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. "High Wage Workers and High Wage Firms." *Econometrica* 67 (2): 251–333.

Airy, George Biddell. 1861. *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*. Macmillan.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Balestra, Pietro, and Marc Nerlove. 1966. "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas." *Econometrica* 34 (3): 585–612.

Bessler, David A., Jeffrey H. Dorfman, Matthew T. Holt, and Jeffrey T. LaFrance. 2010. "Econometric Developments in Agricultural and Resource Economics: The First 100 Years." *American Journal of Agricultural Economics* 92 (2): 571–89.

Bollinger, Christopher R. 2003. "Measurement Error in Human Capital and the Black-White Wage Gap." *Review of Economics and Statistics* 85 (3): 578–85.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2022. "Discretizing Unobserved Heterogeneity." *Econometrica* 90 (2): 625–43.

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped Patterns of Heterogeneity in Panel Data." *Econometrica* 83 (3): 1147–84.

Bronfenbrenner, Martin. 1944. "Production Functions: Cobb-Douglas, Interfirm, Intrafirm." *Econometrica* 12 (1): 35–44.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna. 2024. "Difference-in-Differences with a Continuous Treatment." NBER Working Paper 32117.

Chauvenet, William. 1871. *A Manual of Spherical and Practical Astronomy*. Vol. 2. J.B. Lippincott.

Daniels, Henry E. 1939. "The Estimation of Components of Variance." Supplement to the *Journal of the Royal Statistical Society* 6 (2): 186–97.

de Chaisemartin, Clément, and Xavier d'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.

de Chaisemartin, Clément, and Xavier d'Haultfœuille. 2023. "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey." *Econometrics Journal* 26 (3): C1–30.

Dupont-Kieffer, Ariane, and Alain Pirotte. 2011. "The Early Years of Panel Data Econometrics." *History of Political Economy* 43 (S1): 258–82.

Econometric Society. 1955. "Report of the Montreal Meeting, September 10–13, 1954." *Econometrica* 23 (3): 324–37.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. 2nd ed. Oliver and Boyd.

Fox, Karl A. 1986. "Agricultural Economists as World Leaders in Applied Econometrics, 1917–33." *American Journal of Agricultural Economics* 68 (2): 381–86.

Goldsmith-Pinkham, Paul. 2024. "Tracking the Credibility Revolution across Fields." https://doi.org/10.48550/arXiv.2405.20604.

Griliches, Zvi. 1957. "Specification Bias in Estimates of Production Functions." *Journal of Farm Economics* 39 (1): 8–20.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46 (6): 1251–71.

Heady, Earl O. 1946. "Production Functions from a Random Sample of Farms." *Journal of Farm Economics* 28 (4): 989–1004.

Hildreth, Clifford. 1950. "Combining Cross Section Data and Time Series." Cowles Commission. Unpublished.

Hill, Terrence D., Andrew P. Davis, J. Micah Roos, and Michael T. French. 2020. "Limitations of Fixed-Effects Models for Panel Data." *Sociological Perspectives* 63 (3): 357–69.

Hoch, Irving. 1957. "Estimation of Agricultural Resource Productivities Combining Time Series and Cross-Section Data." PhD diss. University of Chicago.

Hoch, Irving. 1958. "Simultaneous Equation Bias in the Context of the Cobb-Douglas Production Function." *Econometrica* 26 (4): 566–78.

Hoch, Irving. 1962. "Estimation of Production Function Parameters Combining Time-Series and Cross-Section Data." *Econometrica* 30 (1): 34–53.

Imai, Kosuke, and In Song Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for

Causal Inference with Panel Data." *Political Analysis* 29 (3): 405–15.
Jakiela, Pamela. 2021. "Simple Diagnostics for Two-Way Fixed Effects." https://doi.org/10.48550/arXiv.2103.13229.
Kling, Jeffrey R. 2006. "Incarceration Length, Employment, and Earnings." *American Economic Review* 96 (3): 863–76.
Lumsdaine, Robin L., Ryo Okui, and Wendun Wang. 2023. "Estimation of Panel Group Structure Models with Structural Breaks in Group Memberships and Coefficients." *Journal of Econometrics* 233 (1): 45–65.
Marschak, Jacob, and William H. Andrews Jr. 1944. "Random Simultaneous Equations and the Theory of Production." *Econometrica* 12 (3/4): 143–205.
Miller, Douglas L. 2023. "An Introductory Guide to Event Study Models." *Journal of Economic Perspectives* 37 (2): 203–30.
Millimet, Daniel L., and Marc F. Bellemare. 2024. "On the (Mis)Use of the Fixed Effects Estimator." IZA Discussion Paper 16202.
Morgan, Mary S. 1990. *The History of Econometric Ideas.* Cambridge University Press.
Mundlak, Yair. 1961. "Empirical Production Function Free of Management Bias." *Journal of Farm Economics* 43 (1): 44–56.
Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46 (1): 69–85.
Mundlak, Yair. 2011. "Plowing through the Data." Annual Review of Resource Economics 3: 1–19.
Mundlak, Yair, and Irving Hoch. 1965. "Consequences of Alternative Specifications in Estimation of Cobb-Douglas Production Functions." *Econometrica* 33 (4): 814–28.
Nerlove, Marc. 2005. *Essays in Panel Data Econometrics.* Cambridge University Press.
Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49 (6): 1417–26.
Papadopoulos, Alecos. 2022. "The Effects of Management on Production: A Survey of Empirical Studies." In *Handbook of Production Economics*, edited by Subhash C. Ray, Robert G. Chambers, and Subal C. Kumbhakar, 1651–1697. Springer.
Reder, Melvin W. 1943. "An Alternative Interpretation of the Cobb-Douglas Function." *Econometrica* 11 (3/4): 259–64.
Reiss, Franklin J. 1949. "Measuring the Management Factor." *Journal of Farm Economics* 31 (4): 1065–72.
Reiss, Franklin Jacob. 1952. "Individual Differences in Entrepreneurial and Managerial Ability among Illinois Farm Operators." PhD diss. University of Illinois at Urbana-Champaign.
Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58.
Scott, James C. 1999. *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed.* Yale University Press.
Stock, James H., and Francesco Trebbi. 2003. "Retrospectives: Who Invented Instrumental Variable Regression?" *Journal of Economic Perspectives* 17 (3): 177–94.
Sun, Liyang, and Jesse M. Shapiro. 2022. "A Linear Panel Model with Heterogeneous Co-efficients and Variation in Exposure." *Journal of Economic Perspectives* 36 (4): 193–204.
Tintner, Gerhard. 1944. "A Note on the Derivation of Production Functions from Farm Records." *Econometrica* 12 (1): 26–34.
Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data.* MIT Press.
Wooldridge, Jeffrey M. 2021. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators." http://dx.doi.org/10.2139/ssrn.3906345.
Wright, Philip Green. 1928. *The Tariff on Animal and Vegetable Oils.* Macmillan.
Zilberman, David. 2016. "Remembering Yair Mundlak, Scholar and Leader." June 6. https://professorzilberman.com/2016/06/06/remembering-yair-mundlak-scholar-and-leader/.

Journal of Economic Perspectives—Volume 39, Number 2—Spring 2025—Pages 275–282

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by e-mail at <taylort@macalester.edu>, or c/o Journal of Economic Perspectives, Macalester College, 1600 Grand Ave., Saint Paul, MN 55105.

Smorgasbord

Robert Fredona, Sophus A. Reinert, and Teresa da Silva Lopes survey "Forms of Capitalism" (*Business History Review*, Spring 2024, 98:1, pp. 3–35, https://www.cambridge.org/core/journals/business-history-review/article/forms-of-capitalism/03192B7DAF615944FAC5FA7E44A17261). "Simple dictionary definitions aside, after a century and a half of good faith attempts by some of our keenest minds, we don't seem any closer to meaningful agreement about the definition of capitalism or the historical boundaries of the phenomenon. The caution proposed by Weber in defining 'religion'—'definition can be attempted, if at all, only at the conclusion of the study'—should perhaps be applied to studies of 'capitalism' in equal or greater proportion. Some definitions seem too inclusive. N.S.B. Gras's own definition, for example: 'a system of getting a living through the use of capital,' by which he means 'goods or trained abilities used in producing other goods or services.' Even more inclusive is Deirdre McCloskey's. She has argued with panache that capitalism and

■ *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.20251441.

the market economy, which 'contrary to what you might have heard, has existed since the caves,' are synonymous. 'Market participants are capitalists. You are, for example.' Some seem rather too exclusive: despite his protests to the contrary, Braudel's insistence on separating capitalism—'a world apart where an exceptional kind of capitalism goes on, to my mind the only real capitalism'—from both material life and the market economy, and finding it only in the 'shadowy zone' of great merchants and monopolists 'hovering above the sunlit world of the market economy,' seems rigid and artificial (or at least excessively Olympian). . . . Many of capitalism's staunchest defenders tell us that capitalism without competition isn't capitalism at all. Peter Thiel—a capitalist by all the definitions we've read— tells us, no, 'actually capitalism and competition are opposites.' The best definitions risk being somewhat boring. The most interesting ones seem less interested in capitalism's form than its spirit. . . . Many of the newly-coined 'forms' of capitalism that we listed above—perhaps especially the most outré ones, like 'sugar daddy capitalism' or 'Candy Crush capitalism'—may be ways of identifying not new 'forms of capitalism' in the traditional sense but new characteristics of a capacious and polythetically-defined capitalism. Along the same lines, but from a different vantage point, we might think about what the characteristics are that are shared by both 'managerial capitalism' and 'booty capitalism,' or 'mercantile capitalism' and 'casino capitalism.' From either perspective, though, we might put it this way: capitalisms form a family."

Douglas A. Irwin presents "Does Trade Reform Promote Economic Growth? A Review of Recent Evidence" (*World Bank Research Observer*, February 2025, 40:1, pp. 147–184, https://academic.oup.com/wbro/advance-article-abstract/doi/10.1093/wbro/lkae003/7658088). "The findings from recent research, however, have been remarkably consistent. For developing countries that are behind the technological frontier and have significant import restrictions, there appears to be a measurable economic payoff from more liberal trade policies. . . . [A] variety of studies using different measures of policy have found that economic growth is roughly 1.0–1.5 percentage points higher for countries that undertake trade reforms. Several studies suggest that this gain cumulated to about 10–20 percent higher income after a decade. The effect is heterogeneous across countries, because countries differ in the extent of their reforms and the context in which reform took place. Understanding that heterogeneity, which is sometimes attributed to labor market rigidities, financial frictions, or service-sector in puts, merits further research. At a microeconomic level, the gains in industry productivity from reducing tariffs on imported intermediate goods are even more sharply identified. They show up time and again in country after country."

Mario Larch, Serge Shikher, and Yoto V. Yotov discuss "Estimating Gravity Equations: Theory Implications, Econometric Developments, and Practical Recommendations" (Drexel University, LeBow College of Business, Center for Global Policy Analysis Working Paper 25–01, January 3, 2025, https://ideas.repec.org/p/drx/wpaper/2025001.html). "The gravity equation of trade is the most successful empirical model in (international) economics, and hundreds of thousands of academic papers and policy reports have used the gravity model to

estimate the effects of various determinants of trade flows. Proper estimation of the gravity equation will lead to consistent and unbiased estimates of the trade elasticities for various policies. . . . [W]e believe that (apart from the other data) most, if not all, of the developments in the estimating trade gravity literature, are directly applicable to gravity settings beyond international trade, e.g., commuting, migration, FDI, financial-asset flows, cross-border patents, mergers, and acquisitions, etc., . . . We make the following five recommendations within the 'Data' category: (i) Use data on bilateral trade flows for all possible countries; (ii) Use administrative data, on nominal trade flows in common currency, at delivered prices; (iii) Use disaggregated data; (iv) Use panel data for consecutive years; and (v) Include domestic trade data. We also make six additional recommendations on the 'Estimation' of the gravity model: (vi) Estimate gravity in its multiplicative form using PPML [Poisson Pseudo Maximum Likelihood]; (vii) Use exporter-time and importer-time fixed effects; (viii) Employ asymmetric country pair fixed effects; (ix) Carefully model bilateral trade costs; (x) Allow non-discriminatory trade costs; and (xi) Cluster standard errors. Finally, we make four recommendations regarding the 'Heterogeneous' trade costs and policy effects: (xii) Obtain disaggregated policy estimates; (xiii) Allow for dynamic adjustments in the trade costs; (xiv) Consider other types and sources of heterogeneity; and (xv) Consider using heterogeneity-robust DiD [difference-in-difference] methods."

David Chambers, Elroy Dimson, Antti Ilmanen, and Paul Rintamäki discuss "Long-Run Asset Returns" (*Annual Review of Financial Economics*, 2024, 16: 435–458, https://www.annualreviews.org/content/journals/10.1146/annurev-financial-082123-105515). "First, we address one of the central questions in empirical asset pricing, namely, whether stocks consistently beat bonds over the long run. As we noted above, this evidence relies primarily on data series starting in 1900 or, in the case of the United States, 1926, when the University of Chicago's CRSP data starts. In this section, we first discuss the history of stock and bond returns before 1900/1926 . . . The main message is that the US equity premium over (government) bonds at and above 5% in the 1900s was considerably higher than estimates for the 1800s ranging between +1.6% and −0.6%. . . . The 224-year estimate of the annualized UK equity-bond premium is in the range of 2–3%, which is again well below the twentieth-century premium of 4–5%."

Yaqub Chaudhary and Jonnie Penn warn "Beware the Intention Economy: Collection and Commodification of Intent via Large Language Models" (*Harvard Data Science Review*, December 30, 2024, https://hdsr.mitpress.mit.edu/specialissue5). "The rapid proliferation of large language models (LLMs) invites the possibility of a new marketplace for behavioral and psychological data that signals intent. This brief article introduces some initial features of that emerging marketplace. We survey recent efforts by tech executives to position the capture, manipulation, and commodification of human intentionality as a lucrative parallel to—and viable extension of—the now-dominant attention economy, which has bent consumer, civic, and media norms around users' finite attention spans since the 1990s. We call this follow-on the intention economy. We characterize it in two

ways. First, as competition, initially, between established tech players armed with the infrastructural and data capacities needed to vie for first-mover advantage on a new frontier of persuasive technologies. Second, as a commodification of hitherto unreachable levels of explicit and implicit data that signal intent, namely those signals borne of combining (a) hyper-personalized manipulation via LLM-based sycophancy, ingratiation, and emotional infiltration and (b) increasingly detailed categorization of online activity elicited through natural language."

Melissa Kearney and Luke Pardue have edited a collection of six essays in *Strengthening America's Economic Dynamism* (Aspen Economic Strategy Group, December 2024, https://www.economicstrategygroup.org/publication/strengthening-americas-economic-dynamism/). For example, David Deming, Christopher Ong, and Lawrence H. Summers discuss, "Technological Disruption in the US Labor Market." "We measure changes in the structure of the US labor market going back over a century. We find, perhaps surprisingly, that the pace of change has slowed over time. The years spanning 1990 to 2017 were less disruptive than any prior period we measure, going back to 1880. This comparative decline is not because the job market is stable today but rather because past changes were so profound. General-purpose technologies (GPTs) like steam power and electricity dramatically disrupted the twentieth-century labor market, but the changes took place over decades. We argue that AI could be a GPT on the scale of prior disruptive innovations, which means it is likely too early to assess its full impacts. Nonetheless, we present four indications that the pace of labor market change has accelerated recently, possibly due to technological change." The other authors and titles are: Michael R. Strain, "Protectionism is Failing and Wrongheaded: An Evaluation of the Post-2017 Shift toward Trade Wars and Industrial Policy"; Brad Setser, " The Surprising Resilience of Globalization: An Examination of Claims of Economic Fragmentation"; Zachary Liscow, "State Capacity for Building Infrastructure"; Jason Furman "Eight Questions—and Some Answers—On the U.S. Fiscal Situation"; and Jennifer Doleac, "Why Crime Matters, and What to Do About It."

EU Competitiveness

Mario Draghi has authored "The future of European competitiveness," two volumes of analysis and recommendations totaling nearly 400 pages (European Commission, September 2024, https://commission.europa.eu/topics/eu-competitiveness/draghi-report_en). "The EU has set out a range of ambitions—such as achieving high levels of social inclusion, delivering carbon neutrality and increasing geopolitical relevance—which depend on maintaining solid rates of economic growth. However, EU economic growth has been persistently slower than in the US over the past two decades, while China has been rapidly catching up. The EU-US gap in the level of GDP at 2015 prices has gradually widened from slightly more than 15% in 2002 to 30% in 2023 . . . The main driver of these diverging developments has been productivity . . . Slower productivity growth has in turn been

associated with slower income growth and weaker domestic demand in Europe: on a per capita basis, real disposable income has grown almost twice as much in the US as in the EU since 2000. At the same time, three external conditions—in trade, energy and defence—that supported growth in Europe after the end of the Cold War have been fading. First, even as domestic growth slowed . . . [b]etween 2000 and 2019, international trade as a share of GDP rose from 30% to 43% in the EU, whereas in the US it rose from 25% to 26%. . . . However, the multilateral trading order is now in deep crisis and the era of rapid world trade growth looks to have passed . . . Second, as relations normalised with Russia, Europe was able to satisfy its demand for imported energy by procuring ample pipeline gas . . . But this source of relatively cheap energy has now disappeared at huge cost to Europe. The EU has lost more than a year of GDP growth while having to re-direct massive fiscal resources to energy subsidies and building new infrastructure for importing liquefied natural gas. Third, the era of geopolitical stability under US hegemony allowed the EU largely to separate economic policy from security considerations, as well as to use the 'peace dividend' from lower defence spending to support its domestic goals. The geopolitical environment is however now in flux owing to Russia's unwarranted aggression against Ukraine, deteriorating US-China relations and rising instability in Africa, which is a source of many commodities that are critical to the world economy."

The IMF Regional Economic Outlook reports on "Europe: A Recovery Short of Europe's Full Potential" (October 2024, https://www.imf.org/en/Publications/REO/EU/Issues/2024/10/24/regional-economic-outlook-Europe-october-2024). "Europe's productivity gap with the global frontier can be traced back to a more limited market size, capital market constraints, skilled labor shortages, and stalled structural reforms. Firm-data analysis shows that Europe's segmented good and services markets are keeping businesses from becoming larger, spending more on R&D, and exploiting economies of scale. Moreover, fragmented capital markets mean that firms do not draw enough on equity financing. As a result, business dynamics are dampened especially in the services sector where start-ups tend to operate with large intangible capital. . . . There is widespread agreement on the sources of Europe's growth weakness. Recently released expert studies come to a similar conclusion that Europe's low productivity is related to lack of market depth and scale. . . . Remaining barriers are considered to be still substantial and have resulted in less investment and innovation than necessary to accelerate growth and productivity to levels seen in other advanced regions. However, value chain integration has stalled since the last decade . . . and substantial barriers to goods and trade flows remain . . . New IMF analysis finds that in 2020 trade costs within Europe were equivalent to a sizable ad-valorem tariff of 44 percent for the average manufacturing sector compared to 15 percent between US states, and as high as 110 percent in the case of services sectors . . ."

Yann Coatanlem and Oliver Coste focus on European technology firms in their discussion of "Cost of Failure and Competitiveness in Disruptive Innovation" (Institute for Economic Policymaking at Bocconi University, Policy Brief, September 2024,

https://iep.unibocconi.eu/publications/policy-briefs/policy-brief-n24-cost-failure-and-competitiveness-disruptive-innovation). "It is now widely understood that the R&D intensity gap of the European Union against the United States is driven by tech sectors: the United States private R&D in tech is now 6 times higher than in the EU. . . . Leveraging a combination of financial analysis, empirical observations, and limited existing literature, we estimate that restructuring costs (that include much more than severance packages) are approximately 10 times higher in countries with high labor protection, such as in Western Europe, than in countries with low labor protection such as in the United States."

Interviews

Jon Hartley interviews Myron Scholes on "Academic Finance, Black-Scholes Options Pricing, and Regulation" (*Capitalism and Freedom in the 21st Century Podcast*, January 5, 2025, https://capitalismandfreedom.substack.com/p/episode-43-myron-scholes-stanford). On the Black Scholes option pricing model: "The underlying theory was published in the *Journal of Political Economy* with the model or given its assumptions. Now we know that every model has an assumption, every model has an error, every model is an incomplete description of reality. How well does the model do in making predictions? And that's the key. Basically the model has done very well over time. There's a lot of people who say the model doesn't do this, the model doesn't do that, but it does pretty darn great. . . . At the time the Black-Scholes model was published was coincident with the birth of the first listed options trading in the Chicago Board Options Exchange in Chicago. . . . That was in 1973. Then it was the case that there was the old grizzly traders who thought they had the experience from the over-the-counter market and the new young Turks . . . So here's an idea with experience only and intuition versus a model. And the young guys had the model . . . Fisher Black made sheets of paper which talked about the delta and the pricing at different levels of the stock price relative to the exercise price. And they could look at the sheets. And there was a war between the grizzly intuition people and the model people, the young Turks who had no intuition, but they had the model. And in a matter of about six months or so, the young Turks had wiped out the grizzlies, okay, the intuition people."

Joe Walker interviews Eugene Fama on the topic "For Whom is the Market Efficient?" (*The Joe Walker Podcast*, December 31, 2024, https://josephnoelwalker.com/eugene-fama-156/). "Well, for almost everybody, the market is efficient in the sense that they don't have information that's not already built into prices. People who have special information, the market's not efficient for them. So let's say insiders, for example, typically have special information. So as far as they're concerned, this stock is not priced totally efficiently because they have information they know will change the price. But for everybody else, assuming it's efficient, it may be a really good approximation. . . . So if you say, tell me about professional investors, I'll say a very small fraction of them show evidence

of having information that isn't already built into the price. . . . If I go out to the public, alright, the market's efficient for everybody out there. . . . [T]his is what I call the joint hypothesis problem. You can't tell me that prices reflect all available information unless you take a stance on what the price should be. So you have to have some model that tells me, for example, what is risk and what's the relation between risk and expected return. And then we can look at deviations from that and see if the market is efficient. . . . You need a model that tells you how prices get formed. So in the jargon that's called a model of market equilibrium. You need to join that with efficiency, then I can test it in the context of whatever model you tell me is determining prices. . . . [S]o you cannot test market efficiency without a story about risk and return, which is a market equilibrium issue. The reverse is also true. You can't test models of market equilibrium without market efficiency. So these two things are like joined at the hip. They can't be separated. People who do market efficiency, they almost don't exist anymore. Everybody takes it for granted in the academic sphere. It's considered uninteresting to test. But everybody that does market efficiency understands the joint hypothesis problem. But it's not that widely recognized among the people who do asset risk and return models. It's implicitly assumed, but they never make it explicit."

Andrew Peaple interviews "Barry Naughton on the State of the Xi Jinping Economy," subtitled "The economist discusses Beijing's recent stimulus efforts, and the long-term problems building up as China's leader implements his model for the country" (*The Wire China*, January 5, 2025, https://www.thewirechina.com/2025/01/05/barry-naughton-on-the-state-of-the-xi-jinping-economy/). "[N]one of the things that we've seen to prop up demand and keep institutional structures intact have yet involved a substantial resolution of large amounts of debt. He [Xi Jinping] keeps refinancing, kicking the can down the road, injecting some funds into the system to keep anybody from failing, but without resolving any of the problems. That's really a problem, because at a certain point you have to clean up the mess. . . . Japan spent almost a decade trying to painlessly restructure a financial system that had suffered a huge reduction in the value of its assets. It was the fundamental problem that lay behind the so-called 'Lost Decade' in Japan; and now China seems to be repeating some parts of that. . . . China's not really experiencing significant productivity growth. That is astonishing, because if we look at this economy that's implementing all these new technologies, we think, wow, that's gotta produce some kind of explosive growth in productivity. But we don't see it. And it's fundamentally because, for example, China is investing in lots of semiconductor equipment plants that are losing immense amounts of money; it's investing in thousands of miles of high speed rail that go where nobody wants to go. There are just these huge, long run implicit costs from not improving the efficiency of your society. Now, of course, on some level, Xi Jinping is making a gamble that all these technologies will at some point come together and produce a sudden surge of productivity. And he might be right. We can't say for sure that he's not. But thus far, he's very much not."

Discussion Starters

Clark Packard and Scott Lincicome explore "Presidential Tariff Powers and the Need for Reform" (Cato Institute, October 9, 2024, Briefing Paper No. 179, https://www.cato.org/briefing-paper/presidential-tariff-powers-need-reform). "Several US laws provide the president with vast and discretionary authority to unilaterally impose sweeping trade restrictions, and no institution—not Congress, not domestic courts, not US international agreements—provides a quick, sure-fire check on such actions. Thus, while the durable implementation of broad and damaging US tariffs is not guaranteed, its risk—and related economic and geopolitical risks—will remain real and substantial until US law is changed to limit presidential tariff powers. We therefore recommend Congress enact such amendments immediately."

Paolo Santori investigates "Domination vs. Persuasion: The Role of *Libido Dominandi* in Adam Smith's Thought" (*Review of Politics*, 2025, pp. 1–18, https://www.cambridge.org/core/journals/review-of-politics/article/domination-vs-persuasion-the-role-of-libido-dominandi-in-adam-smiths-thought/EA4FCC30F38FC56CDFDE2A17F1976ACD). From the abstract: "Adam Smith argued that human beings naturally desire to dominate others and that they enjoy it. He showed how ancient masters, landlords, and economic actors in some eighteenth-century English and colonial markets were driven by their love of domination against their own economic interests. . . . This article . . . [shows] that, for Smith, the love of domination has nothing to do with the love of praise but that most of the pleasure people derive from it is to see their ends promoted by others without the need to persuade them about the utility of those ends. This understanding locates the love of domination outside commercial society where, under certain socio-economic circumstances, mutual persuasion among individuals is the rule."

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary access to JEP articles, go to the AEA website: http://www.aeaweb.org. Change of address notice must be received at least six weeks prior to the publication month.

The Journal of
Economic Perspectives

Spring 2025, Volume 39, Number 2

Symposia

Drug Pricing and Regulation

Income Inequality

Bond Markets

Features

AMERICAN ECONOMIC ASSOCIATION