The Journal of

Economic Perspectives

A journal of the
American Economic Association

Fall 2022

The Journal of
Economic Perspectives

A journal of the American Economic Association

Editorial offices:
Journal of Economic Perspectives
American Economic Association Publications
2403 Sidney St., #260
Pittsburgh, PA 15203
email: jep@aeapubs.org

The *Journal of Economic Perspectives* gratefully acknowledges the support of Macalester College.
Registered in the US Patent and Trademark Office (®).

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, USA.
Printed by LSC Communications, Owensville, Missouri, 65066, USA.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Summer 2022, Vol. 36, No. 3. The JEP is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. Annual dues for regular membership are $24.00, $34.00, or $44.00 depending on income; for an additional $15.00, you can receive this journal in print. The journal is freely available online. For details and further information on the AEA go to https://www.aeaweb.org/.
Periodicals postage paid at Nashville, TN, and at additional mailing offices.
POSTMASTER: Send address changes to the *Journal of Economic Perspectives*, 2014 Broadway, Suite 305, Nashville, TN 37203. Printed in the U.S.A.

The Journal of
Economic Perspectives

Contents

Volume 36 • Number 4 • Fall 2022

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 3–28

Is There Any Future for a US Labor Movement?

Suresh Naidu

Almost 15 years ago, a previous *Journal of Economic Perspectives* article on American unions (Hirsch 2008) argued that due to increased competition and dynamism in the US economy, the unions created and sustained by the National Labor Relations Act were sclerotic dinosaurs to be replaced, if at all, by new institutions of worker voice that "must flourish in the US economic environment of open, competitive, and dynamic markets." Today, this view of the overall American economic environment looks sanguine; noncompetitive labor and product markets in the last 40 years are now well-documented (Philippon 2019; Naidu, Posner, and Weyl 2018), and the associated rise in inequality, both across workers and between capital and labor, is now a much larger concern.

At the same time, there has been both a resurgence of public interest in unions as well as policy interest from both conservatives and liberals in the United States. Even as private sector union density has continued to decline to around 6 percent of employment, COVID-19 and the subsequent labor shortage generated a spate of prominent examples of collective action among food, retail, and health care workers. I write this paper in fall 2022 during a wave of Starbucks union petitions, with over 5,000 workers having won union elections in the last six months. There is a recent and remarkable win by an independent union at a Staten Island Amazon Warehouse in New York City. This flurry of activity was preceded by "Striketober 2021," with over 100,000 private sector workers (including graduate students at my university) having authorized strike votes, the most in decades. That said, these are

■ *Suresh Naidu is Professor of Economics and International and Public Affairs, Columbia University, New York City, New York. His email address is sn2430@columbia.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.3.

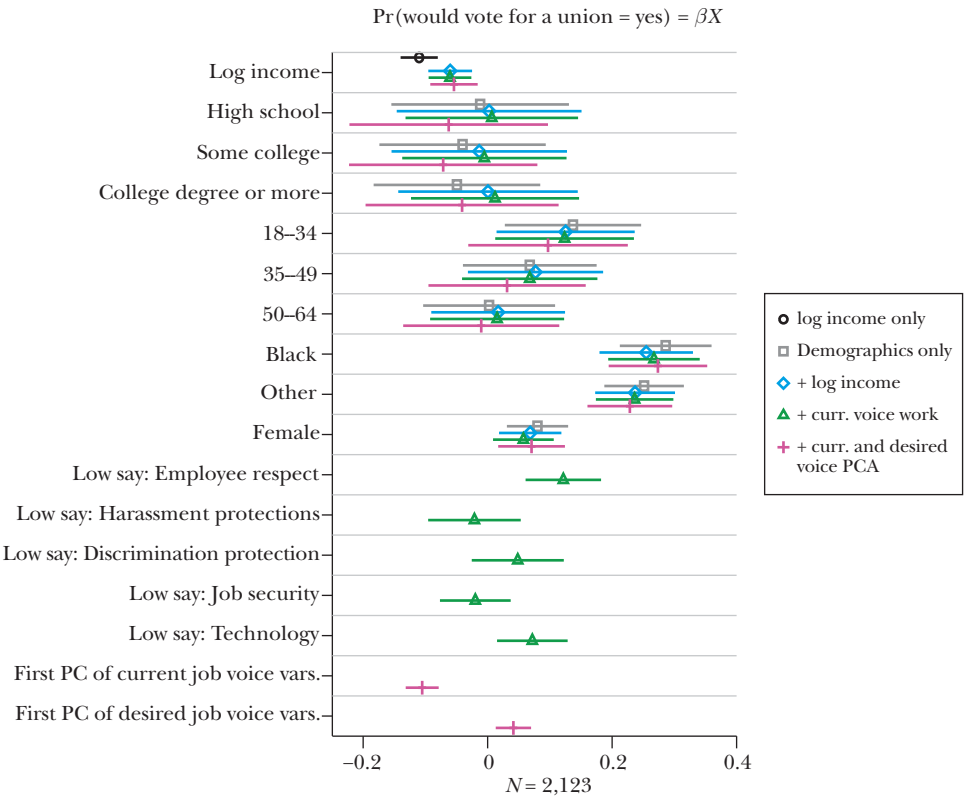
all drops in the bucket of the overall American labor market, and may or may not be harbingers of a resurgent unionism. What future is there for the labor movement in the United States?

A union weaves workplace ties between coworkers together into organizations capable of coordinating collective action sufficient to force a binding bargain with a large firm, a whole sector, or to influence politics. Unions are independent organizations aggregating and representing the interests of workers, and thus have no substitute in the form of government wage mandates or employment regulation. The primary obstacle to widespread unionization in the United States is that labor law and employer opposition requires a high level of workplace social capital to win union recognition, and even more to win a collective bargaining agreement. This slow and costly process has struggled to outpace the exit and downsizing of already-unionized firms. Between the employer-side advantages given by US labor law and diminished workplace social capital, it is difficult to see a path to a persistent increase in union density that is not concomitant with a rewiring of workplace networks and a transformation of American labor law. Nevertheless, COVID-19 and its aftermath may have precipitated the required rewiring, with younger workers transmitting their desire for unions to other segments of the labor market, aided by a sympathetic federal government and exceptionally tight labor markets.

Unions remain very popular. Over 70 percent of Americans approve of unions in recent Gallup polls (McCarthy 2022). Surveys since the 1970s have also asked: "Would you vote for a union if an election were held tomorrow?" Figure 1 shows coefficients from a regression trying to explain variation in responses to this question for nonunion private sector workers, as asked by Hertel-Fernandez, Kimball, and Kochan (2022). Income, race, youth, and gender remain strong predictors of union support, along with those who have experienced low respect (Dube, Naidu, and Reich 2021) and those with low input into technology use in their workplace. Workers experiencing more voice at work are less willing to unionize, while workers who desire more voice are more willing. Yet, as Farber et al. (2021) shows, union density has fallen most in the low end of the unionized educational distribution and among nonwhite workers, arguably the segments with the highest latent demand. This pattern suggests some institutional friction hindering unionization among those with high stated demand. Leading candidate explanations include employer opposition, which can be blunted by policy and market conditions, as well as inherent difficulties in generating the collective action necessary to overcome the barriers imposed by US labor law.

The traditional economic analysis of unions shows that in laissez-faire labor markets, unions are purely distortionary, analogous to a minimum wage or a monopoly pushing employers to hire only insiders, at higher wages, at the expense of outsiders, efficiency, and profits. In models of the labor market with incomplete contracts (Grout 1984; Acemoglu and Pischke 1998) or imperfect competition (Dodini, Salvanes, and Willén 2021; Manning 2013, ch. 12), however, unions can raise both wages and employment, and even improve productivity. Empirically,

Figure 1
Willingness to Vote for a Union by Demographics, Income, and Demand for Amenities



Source: Author's analysis of data from Hertel-Fernandez, Kimball, and Kochan (2022).
Note: Analysis of 2,508 responses to a survey conducted in 2017 from Hertel-Fernandez, Kimball, and Kochan (2022). Mean of the dependent variable is 0.47. Restricted to private sector non-union workers (with > 20 hours of work). Each set of coefficients are from a separate regression each, weighted and with robust 95% confidence intervals shown by the bars, and should be interpreted relative to a constant term not shown. PCA is the standardized first principal component of 16 dimensions of experienced or desired worker voice.

generations of economists have traditionally focused on the wage and employment effects of unions and labor conflict along with productivity and profitability effects. A more recent literature has discussed political effects, internal politics, and policy determinants of unions (Kremer and Olken 2009; Downey, forthcoming; Feigenbaum, Hertel-Fernandez, and Williamson 2018). Comparatively little work in economics has focused on the social networks, workplace conflict, and dynamics of collective action that characterize US labor organizations.

A Background on American Collective Bargaining Institutions

US Unions in Comparative Perspective

Traditional unions bundle two different services: (1) taking wages and working conditions out of the hands of firms and markets and into a collective bargaining process; and (2) building political and economic power by connecting and mobilizing social networks and identities generated via shared experiences of work. In the traditional American unionized industry, these two functions are expressed by the same organization: a labor union that negotiates a legally binding collective bargaining agreement on behalf of the workers covered by it and then bargains over and enforces that agreement using the collective action capacity of its members.[1]

In other countries, these two functions have been disaggregated in different ways. In some, government policies or centralized contracts set wages throughout the distribution, ranging from minimum wages to wage boards to sectoral bargaining and contract extension to nonunion employers. Other countries have also preserved independent membership-based labor organizations, which may provide members with valuable services (for example, unions supply unemployment insurance to their members in traditional "Ghent" systems in a number of European countries) as well as exercise economic and political power with the capacity to strike, educate, and mobilize workers where they work.

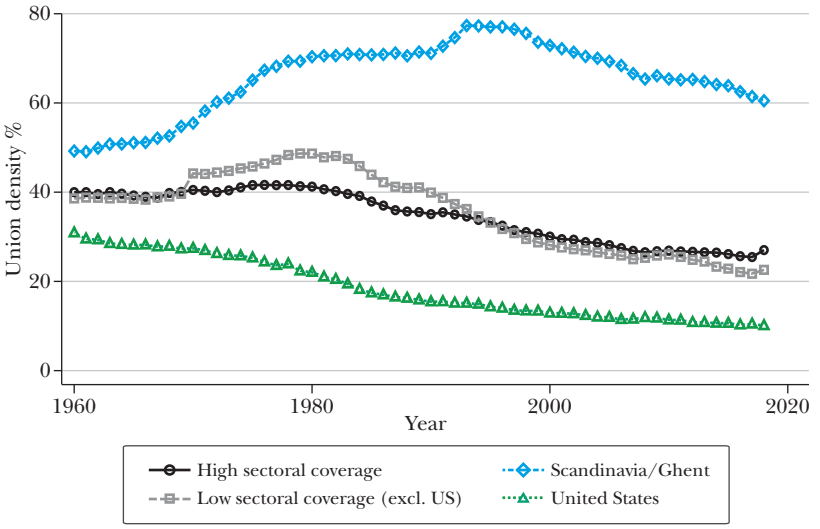
Figure 2 shows patterns of union density across different institutional arrangements in the advanced democracies. Union density has remained highest in countries that have maintained both sectoral coverage of union contracts, where union contracts are extended to all employers in a sector, as well as Ghent-style selective benefits. While some claim that sectoral coverage would make organizing new members easier (for example, Madland 2021), sectoral coverage alone has only preserved union density a little bit more across countries: there is little incentive for nonunion workers not to free-ride on the contract negotiated by the union. But even compared to other countries without sectoral bargaining or any selective benefits, US union density is low, and declining more quickly over the past few decades.

In the United States, private-sector unionization is governed by the 1935 National Labor Relations Act, also called the Wagner Act, subsequently modified by the Labor-Management Relations Act of 1947, also called Taft-Hartley, and the 1959 Labor-Management Reporting and Disclosure Act, also called Landrum-Griffin. The NLRA set up the National Labor Relations Board, which is the primary authority for deciding issues related to union recognition and representation and precludes any state or city from regulating around it. These laws together delineate the process for getting a set of workers legally covered by a union with which an employer has a duty to bargain.

The process of union recognition under US labor law involves a number of steps. First, a set of workers of at least 30 percent of a proposed bargaining unit files

[1] For an overview of the varieties of worker organizing, above and beyond unions, see the review of the landscape from Kochan et al. (2022), published by the Worker Empowerment Research Network.

Figure 2
Cross-Country Union Density



Source: Visser (2019).
Note: High sectoral coverage countries are Austria, France, Germany, Italy, the Netherlands, and Norway. Low sectoral coverage countries are Canada, Ireland, Japan, New Zealand, and the United Kingdom. The Ghent/Scandinavia countries with union insurance are Belgium, Denmark, Finland, and Sweden. The comparison here is restricted to balanced sample of countries. Union density is as a share of employed wage workers as in employment or household surveys. Sectoral coverage means that a union negotiates binding national or regional wage agreements.

cards petitioning for an election, either affiliating with an existing union local or forming a new "independent" union. Second, if an employer has agreed to "card check neutrality," then simply getting a majority of workers to sign cards is sufficient to win a union. Otherwise, the National Labor Relations Board decides whether the proposed bargaining unit is a legitimate "community of interest" and then authorizes a secret ballot election for recognition. Third, both the union and the employer campaign until the election. Illegal tactics are reported as "unfair labor practices" and are adjudicated by the NLRB. Fourth, if the majority of workers vote in favor of a union, the employer has an obligation to bargain with the union in good faith. Fifth, if a collective bargaining agreement is reached (which happens only about half the time), the agreement governs set wages, benefits, and a variety of workplace governance conditions for all the workers covered by it. Sixth, once signed, many contracts are enforced by a grievance procedure, mediated by a worker designated as "steward" who acts as intermediary between workers and their employer. Finally, if no collective bargaining agreement is reached in a year, the union can be decertified via another petition and election.

The National Labor Relations Act protects collective action at the "bargaining unit" level, which is a mix of job categories and geographic establishment. The premise of the law is that the establishment-level bargaining unit is the natural level at which workers share interests, and implicitly, that the barriers to collective action at that level are relatively easy to overcome. However, a hostile legal regime and transformations in employment have invalidated the presumptions on which establishment-level bargaining was built. As a result, the NLRA is as much a legal graveyard as it is a sanctuary for American unions.

In the United States, any worker that wants a union cannot just join one, but instead needs to persuade 50 percent of their coworkers, which means that the decline of US unions is tied up with other forces that have hampered collective action. One reason a variety of new labor organizations, such as the National Domestic Workers Alliance or United for Respect (which seeks to advance the interests of retail workers), have elected to avoid the process as defined by the National Labor Relations Act is that they would become subject to a legal regime that advantages employers, restricts organizational flexibility and tactical innovation, and imposes onerous reporting requirements and regulation. But in doing so, they forfeit the dues revenue that comes with a traditional collective bargaining agreement, and instead rely on philanthropy or other sources of revenue for support.

Strikes as Collective Action

Strikes, the collective withdrawal of labor from an employer or market, remain the reservation position for organized labor in collective bargaining negotiations, and in many countries are also political tools used to pressure governments into policies. While the right to strike is formally guaranteed in virtually every democracy,[2] countries vary in which tactics and degree of coordination they allow workers to engage.

In the United States, the right to strike is technically protected at key junctures in the unionization process, but many of the tactics required to build the collective action and coordination necessary to win strikes are illegal. For example, strikes for union recognition, strikes in response to an unfair labor practice, and strikes during contract negotiation are all protected by the National Labor Relations Act. US law generally allows strikes only at the establishment- or firm-level: specifically, the 1947 Taft-Hartley law forbidding secondary boycotts or political strikes and thus eliminating the possibility that workers in different bargaining units can help each other during labor conflicts. Many tactics to shut down an employer's business, from picketing to workplace occupations, are either extremely circumscribed or illegal under US labor law. Further, beginning in the late 1970s and 1980s, employers began to ignore the prior norm of reinstating striking workers, with ever-increasing use of "permanent replacements" during strikes (Cramton and Tracy 1998). Massenkoff

[2]Notably, a number of "workers' states" restrict the right to strike: China eliminated the legal right to strike in 1982, the Soviet Union de facto abolished it during Stalinism, and communist-run Cuba never granted it.

and Wilmers (2022) show that while strikers experienced higher wages after a strike prior to 1982, since then strikes have resulted in wage losses for workers.

In contrast, many other countries protect broad rights to strike by large groups of workers, sometimes even at the whole economy level. These legal protections do not result in frequent strikes. Instead, strikes are effective as latent, but credible, threats of extremely high costs. A dramatic example showing how European institutions facilitate collective action to regulate employer behavior is provided by the experience of McDonald's coming to Denmark in the 1980s and refusing to pay the union wage negotiated by the hotel and restaurant union. Matt Bruenig (2021) describes the coordinated response by the Danish labor movement:

> In late 1988 and early 1989, the unions decided enough was enough and called sympathy strikes in adjacent industries in order to cripple McDonalds operations. Sixteen different sector unions participated in the sympathy strikes. Dockworkers refused to unload containers that had McDonalds equipment in them. Printers refused to supply printed materials to the stores, such as menus and cups. Construction workers refused to build McDonalds stores and even stopped construction on a store that was already in progress but not yet complete. The typographers union refused to place McDonalds advertisements in publications, which eliminated the company's print advertisement presence. Truckers refused to deliver food and beer to McDonalds. Food and beverage workers that worked at facilities that prepared food for the stores refused to work on McDonalds products.
>
> In addition to wreaking havoc on McDonalds supply chains, the unions engaged in picketing and leaflet campaigns in front of McDonalds locations, urging consumers to boycott the company. Once the sympathy strikes got going, McDonalds folded pretty quickly and decided to start following the hotel and restaurant agreement in 1989.

Even, perhaps especially, in countries with labor peace and low strike rates, the capacity for unions to turn on vast quantities of collective action is the hard power ensuring the soft power of active labor market programs, unemployment benefits, sectoral coverage, and macroeconomic partnership. In contrast, the proscribed strike capacity of US unions since 1947 is perhaps one of the forces driving unions into seeking more political (and even sometimes criminal) sources of power.

Employer Opposition

The typical American employer remains implacably hostile to unions. Even seemingly progressive employers, like Starbucks, major media outlets, and private-sector universities—whose leaders are on the left of the American political spectrum—respond to unionization with the same anti-union law firms and management consultants that less publicly idealistic companies deploy regularly. Over 100 Starbucks workers have been fired seemingly for union activity, and a

number of Starbucks stores that have voted for union recognition have been shut down, all under an executive who was a leading candidate for US Secretary of Labor under Hillary Clinton. Wang and Young (2022) provide credible evidence that the negative employment and survival effects of union wins are driven by managerial opposition to unionization, and other evidence (Dinlersoz, Greenwood, and Hyatt 2017) suggests that this has changed the selection of firms that unions are willing to organize.

Management hostility is not hard to understand. Unions redistribute from capital to labor and reduce the discretion of employers to discriminate in pay (Biasi and Sarsons 2022), to introduce new technologies, and to manage as they see fit. Employers who wish to retain untrammeled authority over their businesses will be averse to the terms of a collective bargaining agreement, which inevitably restricts management's control over the conditions of work, employment, and compensation (Ash, MacLeod, and Naidu 2019).

American labor law gives enormous de facto latitude to employers to fight unions, even as de facto enforcement of labor law and election rules varies with federal administration. Employers can legally use work hours to campaign against the union, union organizers are prohibited access to private worksites, and employers can contest legal definitions of bargaining units and employee status. While firing workers for unionization is technically illegal, there is extreme forbearance towards employers, with the worst punishment the National Labor Relations Board can impose on an employer being a public reading of the law in the workplace.[3]

Unfair labor practices often take too long to adjudicate, and the financial penalties are so small that they pose no deterrent to anti-union activity. Indeed, human resource textbooks sometimes advise managers just to follow certain unfair labor practices as part of the costs of avoiding the union. On top of the lopsided structure of labor law, there is a tactically sophisticated, experienced, and well-funded industry of anti-union consultants (Logan 2006), whose impact on union campaigns deserves further research. Frandsen (2017) finds that unions lose in close elections much more frequently than would be predicted by chance alone,[4] and that this outcome is more likely to occur when Republican appointees are the majority of the National Labor Relations Board.

One reason for the recent upsurge in union activity is that employer opposition has been checked, not by law, but by the historically exceptionally tight labor market. When labor markets are as competitive as they have been in the past two years, the threat of firing does not look nearly as intimidating as in normal times. In

[3] In the case of *Conair v. NLRB,* the US Court of Appeals for the District of Columbia (1983) restrained the most severe penalty the National Labor Relations Board had previously used, which was mandatory bargaining

[4] A common method of looking at the effect of unions on wage premiums and other outcomes, beginning with DiNardo and Lee (2004), is to compare companies where a union barely won an election to companies where the union barely lost. This regression discontinuity design assumes that companies just above and just below these thresholds are valid comparisons. But the finding that at some times the outcomes of close union elections are asymmetric calls this research design into question in this setting.

this way the demand for voice in the workplace may be significantly complemented by the opportunities for exit from an existing job.

Collective Action and Social Networks at Work

The structure of US labor market institutions makes the level of workplace collective action necessary to win recognition and a contract higher and harder to overcome than in most other advanced democratic countries. It puts a particular onus on the "bargaining units" to withstand the hostility of an employer with a clear interest in preventing unionization. Scholars have pointed out numerous reasons why workplace collective action is difficult: the traditional free-rider problem (Olson 1965); the diversity of interests inherent in workers who are selected by employers for synergies in production, rather than shared interests (Offe and Wiesenthal 1985); and the high turnover for some groups of workers coexisting with high employer loyalty of others (Reich and Bearman 2018). Together, these forces erode the "social capital at work" and the associated workplace social networks.

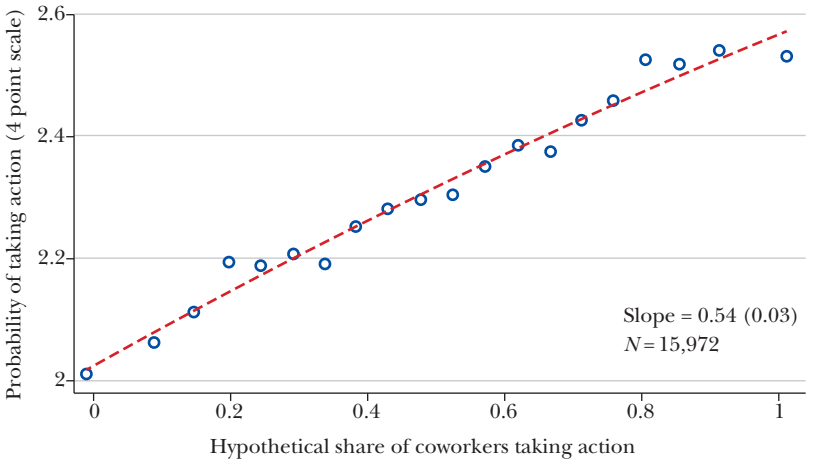
One reason that social networks at work matter is the presence of strategic complementarities in workplace collective action. Figure 3 shows the results from conjoint experiments run during 2020 (for details, see Hertel-Fernandez et al. 2020). We asked workers how willing they were to do a variety of collective actions if p percent of their coworkers were willing to do them, where p was hypothetically randomized. We found that workers were more willing to undertake a given form of collective action when a larger fraction of their coworkers were doing it as well, which suggests that strategic complementarities (not strategic substitutes, as in classical free-rider problems) are pervasive in worker collective action. Coordination seems to be the obstacle to collective action revealed by these data (although anecdotally, free-riding is also pervasive!).

Another reason that social networks at work matter is because of social learning about union advantages and disadvantages from coworkers, as in classic models of network learning. Unlike most models of network learning, there is an important component of secrecy involved, because once employers get wind of an organizing campaign, a tremendous amount of counter-union persuasion, often targeting the same central workplace leaders, begins.

Strategic interactions and social learning create important roles for network structure among coworkers and co-union members (Ballester, Calvo-Armengol, and Zenou 2006; Galeotti, Golub, and Goyal 2020), in particular the centrality of pro-union workers. In the labor organizing context, there is an explicit folk wisdom around the importance of targeting "workplace leaders" for persuasion (McAlevey 2016). Activist workers and union organizers rely on features of workplace social networks to persuade people.[5]

[5] In broader social movements, experimental and quasi-experimental research by Bursztyn et al. (2021) and González (2020) show the importance of social networks and relationships in generating collective action in Hong Kong and Chile, respectively.

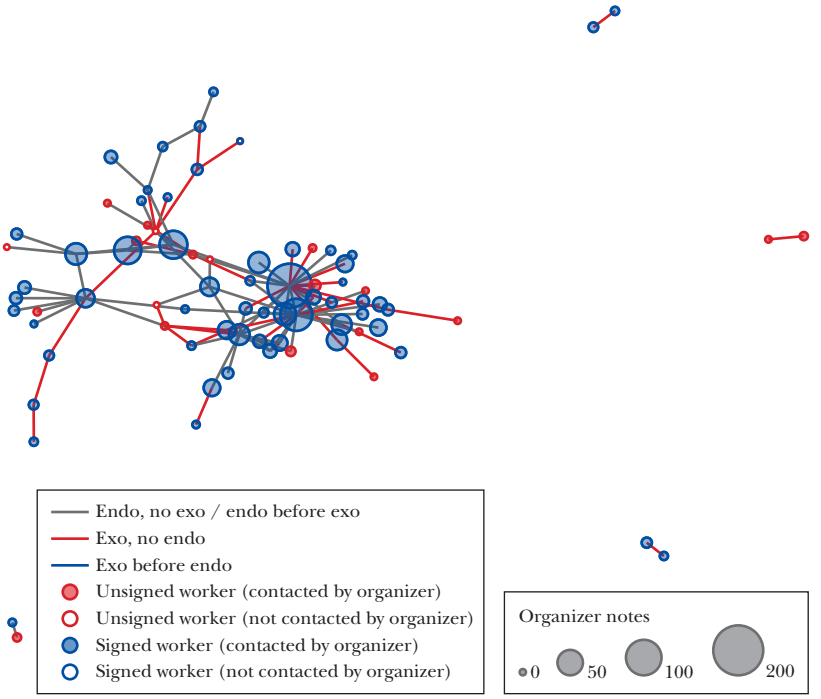
Figure 3
Strategic Complementarity in Worker Collective Action



Source: Author's analysis of data from Hertel-Fernandez et al. (2020).
Note: The horizontal axis shows the survey respondent willingness to participate in a given type of collective action (protest, strike, pay dues to organization, start a union, sign a petition) as function of hypothetical share of co-workers willing to participate in same action. Slope estimates from the specification $Willingness_{ij} = \beta Share\,Coworkers_{ij} + \delta_j + \delta_i + \xi_{ij}$ where j denotes type of action and i denotes respondent. Data is described in more detail in Hertel-Fernandez et al. (2020). The specific question text is: "How likely would you be to [sign a petition/start a union/pay dues to an organization/participate in strike/participate in a protest] to improve working conditions at your employer if you know that p percent of co-workers were willing to do the same action. By coworkers, we mean all others at your workplace who are not managers or supervisors."

Evidence that these workplace social networks matter can be found in Shepherd et al. (2022), where we look at how organizing success, measured by cards signed, varies with how extensively organizers use the network information at their disposal. We create a measure of "network-driven organizing" across 121 retail organizing campaigns, by machine-processing organizer field notes and by measuring the correlation of organizer attention (measured as share of field notes) with worker centrality in the network map of workers (recovered from the co-mentions of workers in the field notes). Figure 4 shows an example of highly network-driven organizing store, along with the cards signed by each worker and indicators for whether the link was created by the organizer (exogenous) or existed independently (endogenous). We find that when the correlation between organizer attention and worker centrality is 0, the number of cards per campaign worker-week is almost 40 percent lower than when the correlation is 1. The low base rate of cards signed, with only 20 percent of workers ever signing in a store, also suggests the structural problem: while organizer strategy matters, changes in organizer strategy alone are unlikely to get to a majority of worker support.

Figure 4
Workplace Networks and Labor Organizing Outcomes



Source: Based on data from Shepherd et al. (2022).
Note: Social networks and labor organizing success at a specific organizing campaign. Edges represent social ties as recovered from the organizer's field notes. "Exo" edges are those edges created by the organizer themselves during the campaign, "Endo" edges are those that existed prior to the campaign. Size of the node indicates the amount of organizer effort in persuasion spent on that worker, and the color indicates whether the worker signed or not, while hollow vs. filled dot indicates whether they had direct contact with the organizer.

Networks matter for sustaining norms of solidarity that facilitate collective action as well. Abstract solidarity, distinct from reciprocity or altruism, is maintained by networks of interaction between workers at work and even outside of work in shared neighborhoods, religious, ethnic, and linguistic groups. Social activities and regular contact serve as glue in allowing high levels of collective action to be sustained. Labor history is filled with accounts of workers punishing norm-breakers, sometimes with ostracism, other times with violence—but these threats also work to sustain high levels of cooperation and collective action. As a historical example provided by Gould (1995), workers in 1848 Paris were organized into crafts with high within-craft interactions and low cross-craft interactions, and so strikes varied by craft during the 1848 revolution. But workers in 1871 Paris were organized on the basis of neighborhoods, so the patterns of collective action in the Paris commune varied more at the neighborhood level.

The importance of networks shows that aggregate measures like union member density do not measure union power accurately. Union density alone does not capture the role of pro-union worker-activists, who are important sources of bridging and bonding social capital in unions. Anchoring union collective action is a set of people who are extremely attached and loyal to their coworkers and the labor movement and who work to strengthen unions despite little in the way of personal benefits. This "militant minority" (Uetricht and Eidlin 2019) forms crucial ties between unions and their (potential) members, constituting shop stewards, canvassers, workplace council members, health and safety reps in union shops, and "salts" (that is, workers who get a job with an intention of organizing the other workers) in non-union shops.

Recent experimental fieldwork by Boudreau et al. (2021) in Myanmar shows the importance of these self-selected union leaders in convincing workers about union wage proposals. Leaders are more altruistic and conscientious than other union members, for example, and groups of workers treated with union leaders are more engaged and more likely to come to consensus on minimum wage proposals closer to the union proposals. Finally, workers are more likely to complete a cost-of-living survey that will help inform the minimum wage when invited by a union leader who is also inviting many other workers, again suggesting union leaders play important roles in networks.

A Decline of Social Capital at Work?

Setting aside the changes in labor law and employer opposition, why might the capacity for worker collective action have declined? One tentative hypothesis is a decline of social capital created at work as a part of a general decline in social capital, particularly among low-education workers.[6] While convincing evidence of this hypothesis must wait, some suggestive evidence can be found in the General Social Survey data on the share of friends who are coworkers, for the group of private-sector workers with a high school education or less declined from 21.5 to 16.4 percent between 1986 to 2002, while it increased from 17.8 to 19.2 percent for those with more than high school education. The Social Capital Project (2017) published by the Joint Economic Committee writes, based on data from the American Time-Use Survey: "Between the mid-1970s (1975–1976) and 2012, the average amount of time Americans between the ages of 25 and 54 spent with their coworkers outside the workplace fell from about two-and-a-half hours per week to just under one hour." Union decline might be seen as yet another form of associational life that has declined for all the same reasons other forms have declined. In this sense, the decline of unions may be as akin to the decline of churches as the decline of heavy manufacturing.

Alongside a decline in social capital could be a decline in work as a source of identity, meaning and dignity in the lives of noncollege workers (Kaplan and Schulhofer-Wohl 2018). In the 1982 General Social Survey, less-than-college-educated

[6] Putnam (2000) writes that "the balance of evidence speaks against the hopeful hypothesis that American social capital has not disappeared but simply moved into the workplace."

workers were more likely to report that their occupation was not recognized or respected. From 2002 to 2014, the survey occasionally asked if respondents were "treated with respect at work," with less-than-college-educated workers reporting significantly less respect at work.

A literature in organizational behavior and personnel economics has examined the role of social capital at work, but has generally focused on its positive effects on productivity and incentives (Bandiera, Barankay, and Rasul 2008; Krackhardt 1990). But some profit-increasing workplace practices may reduce social capital. For example, erratic or even predictable but 24-hour scheduling reduces the ability of workers to coordinate leisure time together, and high-turnover workplaces naturally will find it difficult for a sense of workplace community to develop. Performance pay may increase cross-worker inequality, reducing a sense of shared identity. Beyond these economic determinants, employers might underinvest in developing social capital at work—particularly for workers who have shared interests not shared by an employer—to mitigate the capacity for collective action. One extreme version of reducing social capital at work is "divide and rule": deliberately hiring a linguistically or ethnically heterogeneous workforce in order to prevent collective action (Ferguson 2016).

Social Networks at Work Can Be Built by Organizing

Workplace social networks are not static, and a resurgent labor movement would transform them. Activism and labor actions themselves construct resilient social ties, as argued by many labor ethnographies (Fantasia 1989; Kornblum 1974). Multi-employer union locals can bridge workers across firms, with social and political activities that bring workers together even off-work. In turn, there could be a self-fulfilling labor "quiescence-trap." An energized union holding many actions and constantly involving its membership in group activities can generate relationships among members that may make further collective action easier; in contrast, a bureaucratic, service-oriented union that interacts in a purely transactional way with its membership may find only weak ties among workers when it comes to mobilizing them for collective action.

One role of union organizers is to build autonomous social capital at work and mobilize it towards collective action, beginning with small public actions like petitions and button-wearing, and culminating in an organizing drive and a successful election win certified by the National Labor Relations Board. Rather than taking workplace networks as fixed, union activists and organizers also can create network ties themselves, catalyzing conversations between pro-union workers and other, more noncommittal employees. Building this social capital at work is hard. Unions are fundamentally different from other voluntary organizations exactly because they are organizations defined by firms and labor market boundaries, not voluntary clubs of shared interests. The sorting and self-segregation that may induce strong identities in other voluntary organizations might be muted in unions exactly because most workers do not choose their coworkers: their employers do. The diversity of worker identities and interests

arguably makes workplaces harder to aggregate and organize than other social groups.

Further, the increased political polarization of Americans means that unions, as a key constituency of the Democratic party, immediately lose prospective Republican members. Labor organizing can thus split even harmonious workplaces into politicized factions. A veteran organizer named Jim Straub put it like this (as quoted by Nolan 2020):

> It is truly not just the unfair playing field, or the power of the boss's fight to scare people, that prevents a majority of a workplace from voting to unionize. In many many workplaces, skepticism and disinterest in doing a collective fight thing is widespread, organic and real among the majority in the middle. Not among social science adjuncts, or journalists, or in large urban service job clusters where almost all the workers are poor and nonwhite. In those types of workplaces, I think any competent organizing program should be able to grow the union. But in places that reflect the educational or political diversity of the country as a whole, I think you're working with fewer total supporters and that's why you wind up chasing stuff like card check neutrality.

A form of Baumol's "cost disease of the service sector" afflicts the union organizing process (for discussion, see Baumol 2012). Persuading coworkers and sharing credible information in workplace networks, and doing so covertly, takes time and energy. Labor organizing is a tough job, good organizers are rare, and most people who would be good organizers are also good at other things that pay more and are less demanding. One paradox of reduced discrimination and misallocation of labor may in fact be a weakening of the activist core that made unions successful. If workers who are unusually charismatic and talented were natural leaders in labor movements, more meritocratic hiring and identification of talent (including declining discrimination in race and gender categories, as in Hsieh et al. 2019) may weaken the capacity for collective action in those workers who remain. This change in the composition of workers would generate the observed pattern that the only workers able to benefit from collective action are those that are already relatively skilled (either informally or formally). Thus, the increased returns to interpersonal skills (Deming 2017) may further weaken unions, as the social skills that are increasingly rewarded by the market are now less relatively valuable in labor organizing.

At the same time, technological and organizational changes that encourage workers to interact and socialize with each other, even remotely, can also raise the productivity of organizing effort and rebuild the capacity for collective action. Technological proposals along these lines abound. Workplace communication tools have been incubators for labor organizing both inside and outside the tech sector (for example, as reported by Lawrence and Kramer 2021). Initiatives like Coworker.org allow employees dispersed across the world to sign petitions and discuss workplace issues online, assisting workers interested in collective action with

internal governance and mediation. Online communities can also generate material support. Even so, while the internet is good at having people share information, identity and resources, it has not yet been able to replace the relationships built by shared experiences of work (Blanc 2022). Perhaps there is some future of online labor movements abetted by mechanism designs and online tools that facilitate collective action,[7] but such scenarios remain far away from the vast majority of the labor force today.

If workplaces are no longer places where salient identities are incubated, can other identities be deployed to organize workplaces? The contribution of immigrant workers to the growth of the labor movement in many states is deserving of more quantitative study; immigrants from particular countries clustered in particular sectors and employers, and their tight-knit communities (and often shared experiences of politics back home) were an important resource in increasing union density in large coastal cities (Milkman 2006).[8]

Making Unions Easier to Establish

Reducing employer opposition to unions is another key ingredient to rebuilding union density. Doing this systematically would entail a radical rewrite of US labor law to alter the incentives employers have to evade or oppose unionization. Some possible policies include: (1) sectoral bargaining, paired with mechanisms to incentivize participation in democratic labor organizations, (2) lowering incentives of non-union employers to oppose unions, using both the carrots of pro-union procurement policy and the stick of steep penalties for anti-union activities, and (3) ensuring unions are adding value, at least to particular "high-road" employers, as well as redistributing. With firms as big and as powerful as they are today, it is difficult to see how large increases in private sector union density can happen without government policy to reduce the profitability for firms of opposing unions. But it is also difficult to see how government policy can be adequately reformed without a resurgent labor movement demanding it.

In the United States and other establishment-level bargaining systems, a basic constraint on union density is that it is hard to organize new firms fast enough to keep pace with the exit of already unionized firms. Even if unionization were an order of magnitude easier, the costly trench warfare of establishment-by-establishment organizing in the face of structural change and natural business dynamism makes keeping union density constant, let alone expanding it, an uphill battle. The difficulty of maintaining union density in the establishment-based system is the primary reason why sectoral coverage has emerged as a key demand of labor

[7] As one example, Kellogg tried to recruit permanent replacements online during a recent strike, but people on Reddit r/antiwork filled the application interface with spoofs until it shut down (as described in Thalen 2021).

[8] See also "Bargaining for the Common Good," where labor organizations partner with other groups (for example, climate or racial justice groups) to make broader nonlabor demands on employers as well as making labor demands on governments. For details, see https://smlr.rutgers.edu/faculty-research-engagement/center-innovation-worker-organization-ciwo/bargaining-common-good.

movements: for example, California has recently enacted a law (AB257) for a council to establish minimum standards for wages, hours, and working conditions in the fast-food industry. Sectoral coverage ensures that the hamster wheel of organizing every new shop is turned off, because new employers are automatically covered by the pre-existing agreement, and incumbent employers do not have to worry about cost competition from lower-wage employers in their sector.[9]

Wage Mandates and Sectoral Bargaining

The simplest policy tool for mitigating the incentives for firms to fight unionization is to take labor standards out of competition by legislative action. Thus, higher minimum wages and employment regulations that bind on even nonunion employers are effectively a pro-union policy. Service-sector unions have acknowledged this complementarity and have become active agitators for minimum wage campaigns, driving the "Fight for $15" movement in the United States and generating political pressure for large employers, like Wal-Mart and Amazon, to raise corporate minimum wages significantly.[10] These policies do not directly involve or empower unions.

A more direct tool for taking wages out of competition between union and non-union firms is sectoral coverage or sectoral bargaining, where union-negotiated bargains are automatically extended to all workers in an administratively defined sector (Madland 2021). Historically, these sectoral arrangements are not unknown in the United States; during World War II, the National War Labor Board formed commissions to set wages, and industry-level minimum wages have been set in many states via commissions (Andrias 2019). Unions, when sufficiently powerful, have negotiated "master contracts" or "pattern bargaining" where similar contract terms cover many different establishments. Even when unions are not present, administrative "wage boards" can set wages in tripartite consultation with workers, employers, and government officials, perhaps even raising interest in forming employer associations and worker organizations to manage the process.

Sectoral bargaining is particularly attractive for unions, because it creates a venue for unions to participate in wage- and standard-setting while mitigating employer willingness and ability to avoid unionization. It counters many of the problems facing unions in establishment-level systems, such as a preponderance of small employers, fragmented contracts that each need to be serviced, widespread outsourcing, the National Labor Relations Board recognition process, and high establishment churn. Many inside the labor movement forcefully advocate for

[9] Farber and Western (2001) compute the "steady-state new organizing rate" it would take to maintain union density in the establishment system, finding that "new-organization rate would have to increase by over six times (from 0.09 percent to 0.65 percent) to yield a steady-state union membership rate of 12.25 percent. But this would require that the unions organize each year new members equal to 7.5 percent of their current membership." They conclude, and I agree, that a six-fold increase in the new organizing rate is infeasible.

[10] Clemens and Strain (2020) look at why unions support minimum wage laws and find that union membership increases with minimum wage increases.

sectoral bargaining in the belief that it would address many of the pathologies of the establishment-level bargaining system created by the 1935 Wagner Act.

One limitation of sectoral bargaining proposals is that they do not build unions as independent organizations that can advocate for worker's interests on a broad variety of issues beyond wages. They do not provide any additional incentives for people to join unions, or even for unions to collect additional revenue on behalf of, or remain democratically accountable to, covered workers. In some states, unions have effectively won sectoral bargaining in some key health worker sectors: for example, all workers in home health care in California and Minnesota are covered by contracts negotiated by the Service Employees International Union. These contracts have been won legislatively and do not result in any automatic union membership, as the US Supreme Court (2014) ruling in *Harris v. Quinn* implies that there is no need for covered government workers to even pay "fair share fees" to the union. Thus, these unions still need to expend resources in convincing home health workers to join and pay dues. Some unions have found creative and successful ways to do so, like demanding an opportunity to advertise the union to new workers during orientation and training, but they still have nowhere near a majority of covered workers as members. Beyond the independent importance of unions as organizations, there is also a worry about the sustainability of unions in a sectoral bargaining regime as they continue to dwindle in resources and membership. Who represents workers on sector-wide wage boards when there are no unions?

One could imagine pairing sectoral bargaining proposals with devices for workers to contribute to a representative organization of their choice, with a variety of matching fund mechanisms from the government provided to mitigate free-riding problems. Unions active on sectoral wage boards might also fund their activities by taking on enforcement of sectoral agreements and employment standards more broadly. Unions in several existing sectoral standard associations, such as the role placed by the Service Employees International Union and UNITE-HERE in setting standards for all establishments in airports or Los Angeles hotels, and the Teamsters in setting standards for Seattle taxi drivers, see part of their role as enforcing the sectoral agreement, given widespread employer violation and limited government enforcement resources (Jacobs, Smith, and McBride 2021). There could even be additional government funding for labor organizations as part of enforcement of labor law, taking advantage of the tacit knowledge workers have about workplace conditions, but this doesn't generate the fiscal accountability that dues-paying membership does (Fine 2017).

Making Employer Opposition Costly

Another option is to raise the costs to firms of opposing unions. The Protecting the Right to Organize act, currently stalled in Congress but supported by much of the labor movement and a number of progressive politicians, is a proposal for major reform of the National Labor Relations Act designed to make it harder for employers to resist union efforts. Provisions include increased fines and personal liability for employers who violate labor law, reinstating fired pro-union workers

while their case is pending, and creating a right for workers to sue employers that violate their unionization rights under the NLRA. It also repeals various provisions from the 1947 Taft-Hartley legislation; for example, the rules outlawing cross-sector solidarity strikes, allowing state-level "right-to-work" laws under which worker do not need to join an established union in their place of work, and requiring arbitration and mediation to facilitate first contracts after a union has been established.

Besides penalizing union opposition, government policy can incentivize union recognition by firms. Historically, as with the National War Labor Board during World War II, government procurement policy has encouraged firms to accept unionization. While there are legal difficulties because of potential conflicts that could arise with the National Labor Relations Act, federal spending bills could prioritize unionized firms, or encourage union neutrality in government suppliers or private-public ventures, much as the federal Equal Employment Opportunity Commission does with racial and gender discrimination. For example, the Inflation Reduction Act of 2022 has apprenticeship requirements for firms building some of the new energy facilities eligible for federal subsidies, thus incentivizing contractors to use union construction labor. The White House Task Force on Worker Organizing and Empowerment offers a large list of federal reforms that would facilitate labor organizing, including ideas like this, at https://www.dol.gov/general/labortaskforce.

Going further, one could imagine (in a post-National-Labor-Relations-Board-preemption world) attaching union neutrality provisions to a variety of protections that the government extends to employers, like limited liability. As a far-fetched example, imagine intellectual property law that had collective bargaining rights attached to it: if an employer wants its patent or copyright protected, it has to respect union rights (for example, it cannot be in violation of the National Labor Relations Act). Such requirements would have the satisfying symmetry of pairing a government-granted monopoly with government-ensured labor protections. Patents are monopoly property rights explicitly protected by the government; that protection could be contingent on employers that hold those patents (and their licensees) respecting the free association rights of their workers.

Policy can also reduce the ease with which employers can protect profits from collective bargaining, decreasing the ability of employers to move (or threaten to move) production out of unionized job jurisdictions via outsourcing or subcontracting. While more work is needed to design such policies, a modern-day labor bargaining regime could account for the full value chain, giving workers a chance to bargain over the value otherwise ultimately paid out to managers and owners, regardless of the multiple legal organizations lying in-between.

Productivity-Increasing Unionization

A final device for encouraging employers to accept unions is to ensure that unions add value, rather than simply redistributing. Evidence on the productivity effects of unions exists, but is scarce in the recent literature. Barth, Bryson, and Dale-Olsen (2020) use variation in taxation of union dues in Norway and show that in firms where more workers join unions, productivity and wages both increase. One

reason to expect larger positive productivity effects in the service economy, however, is that there may be much stronger complementarities between unions and service quality. A slogan of many teacher unions was that "teachers' working conditions are students' learning conditions," and teachers' unions have successfully mobilized parent support by demanding workplace conditions that also improve (or are seen to improve) schooling conditions (Hertel-Fernandez, Naidu, and Reich 2021). Many of the determinants of bad working conditions may also be determinants of poor service quality. Using a regression discontinuity design, Sojourner et al. (2015) find that nursing home quality is delivered at lower costs in unionized nursing homes. Naidu and Reich (2018) for example, find that customer Yelp ratings are negatively associated with labor conflict (as measured by cards signed by workers) at Wal-Mart stores. Further, the nonroutine, quasi-specific nature of many service sector jobs may result in training being optimally provided by intermediaries that are not captured by employers, like union apprenticeship programs (Naidu and Sojourner 2020).

Of course, unions can also reduce productivity and firm investment, transferring value from consumers, employers, and outside-the-union workers to inside-the-union workers, given that the union is a democratic organization accountable to the latter and not any of the former. On these grounds, more broadly encompassing unions, as stressed by Calmfors and Driffill (1988) in the context of inflation, internalize many of these externalities on other parties and may be economically preferred to narrow, fragmented unions.

Increasing Demand for Unions among Those Able to Win Certification

In the absence of changes in labor law, employer opposition, and capacities for collective action, increases in American unionization will be driven by increases in demand for unions and collective action among those currently in the best position to win union certification. Unionization might be easiest among those who already have some degree of security, who are employed at firms that have substantial rents to be redistributed, and who already have the social capital at work to address their collective action problems. Increasing demand for unions among such a group can come from the possibility of higher wages, improved job protections, and the sense of dignity and freedom that can come from having a say over the technologies and conditions of work.

Low-Wage Younger Workers

As the college wage premium has fallen, particularly when considered net of tuition and student debt, a number of younger workers have begun looking to unions as possible solutions to dim job-market prospects. These union demands can percolate through low-wage workplaces to infect other low-wage workers as well, often of very different ages and class backgrounds. This percolation may have

been particularly strong among essential workers during COVID-19, many of whom wound up relying on each other, and disappointed in their employers, to a much greater extent than normal. Whether this contagion happens, and happens fast enough to overwhelm employer opposition, is one of the key questions for the American labor movement of the moment.

These types of workers have been successful recently at large employers close to the top of the job ladder in these sectors (for example, in some Amazon warehouses and Starbucks coffee shops). These are exactly the employers with rents to be claimed, as well as where exit is relatively unattractive, even in a tight labor market. The combination of (1) potential gains from collective "voice," (2) relatively low gains from individual "exit," and (3) low costs of unemployment from employer retaliation may together help explain the patterns of new union interest we are seeing today.

Workers in the Knowledge Economy

At the other end of the labor market, an increasing share of labor is deployed in the "ideas sector," including universities, media, and technology. There has been perhaps surprising interest in unionization in these sectors as well. For these workers, within-firm or within-industry wage inequality may be particularly high, creating a demand for union wage compression. In addition, unions or other labor market organization could bargain over intellectual property, publication rights, which customers to serve, political representation, and general issues of voice and other amenities at work. The experience of tech workers organizing with Communications Workers of America, new media affiliations with NewsGuild, and graduate student unionization with the United Auto Workers may be evidence of this margin in action.

While labor organizing in this area may be less of a force for economic equality, these unions might still be important for protecting non-wage amenities (like tenure) that facilitate basic research and govern the allocation of innovative labor. Labor market distortions are rife in the knowledge economy. For example, Goolsbee and Syverson (2019) show that research academics are subject to considerable monopsony power from their employers; Marx (2011) shows that noncompete contracts significantly lower mobility of technical professionals; and Kline et al. (2019) show that innovative firms share rents from patents only with senior employees, all of which suggest pervasive labor market imperfections. The conditions of postdoctoral researchers, for example, most of which do not become tenure-track jobs, could be ameliorated by collective bargaining at universities. Collective bargaining's comeback could be among those workers who expect autonomy and creative freedom as part of their jobs, but do not trust their employers to guarantee it.

Reflecting the "Brahmin left" tendency among the highly educated (Gethin, Martínez-Toledano, and Piketty 2022), the younger workers currently in these industries are more partisan Democrats, on average, than older generations of union members, and an influx of them into the labor movement could alter the internal balances of power and increase within-union political competition. For example,

a recent referendum by the United Auto Workers for direct voting on leadership, as a device to curb corruption among union leaders, was partly swung by graduate student locals.

It remains to be seen if any of these new sparks will result in durable collective bargaining agreements and independent organizations and whether they will spread. If the labor movement grows only via college-educated workers, it will stay small, and the resulting unions will look different from traditional unions. Issues of workplace discrimination, sexual harassment, working from home, surveillance and privacy at work and on social media may become subjects of bargaining. But a labor movement of this sort will still aim to raise wages, secure health care, and compress compensation inequality (and maybe remote work inequality) across workers, and as such may spark demands for unionization among a broader cross-section of workers. While surely not sufficient for regenerating a labor movement, the spate of activity seen in the past year is equally surely necessary.

Conclusion

Some public policies have offered partial substitutes for the wage-setting, workplace health and safety regulation, and collective action roles of unions. Employment law and wage mandates have regulated wages and many workplace characteristics. In many firms, human resources managers solicit feedback on workplace characteristics via surveys and exit interviews. Particularly post-COVID, workplaces are not the fixed-capital-intensive places of the mid-twentieth century, labor turnover is high, and within-firm job differentiation and ideological polarization is higher, all of which diminish the possibility of collective action at work. The cocktail of circumstances—capital-intense manufacturing, workplace-based communities, pro-worker ideology, and extensive public procurement—that gave rise to twentieth-century American unions may not appear again.

Or it might. So long as work occupies such a large share of time for so many people, the process of joint production can generate a set of unique social ties. These networks can be enlisted by employers for their own political or social ends (Hertel-Fernandez 2018), or deployed to facilitate collective action by workers themselves. The gig economy, which may at first seem to separate workers, may paradoxically provide the scaffolding for such an organization: when workers all interact online, the emergence of online fora to coordinate and make demands can be successful. Traditional unions were born in factories that brought together workers who has previously been dispersed in the "putting-out system" (Marglin 1974). Modern platforms centralize jobs that were once too dispersed and marginal to organize, and thus give unions and workers a single organizational target: Jin, Kominers, and Shroff (2021) offer an overview of what unions could look like in platform sectors. The increasing need for caring labor, be it health care, counseling, education, or mental health, will not soon succumb to automation, and indeed is very likely to continue to be subsidized by the government, creating scope for a rise in readily unionized public

employment. Finally, the steady increase in income inequality, and general support for pre-tax measures to curb it (Kuziemko, Marx, and Naidu 2022), will keep labor organizations in the minds of policymakers and advocates.

Rapid increases in union density are like wildfires (or pandemic waves), and I have little confidence in predictions about whether worker organizations will grow, or even persist, in the twenty-first century. If they do, I suspect they will be very different from the labor organizations of the twentieth century. These new organizations, possibly incubated inside or alongside existing labor unions, will depend on government in new and multiple ways, deploy collective action at multiple scales for both economic and political goals, and use and bargain over technology in ways that are hard for any middle-aged academic to anticipate. In the current lopsided legal environment, labor market tightness has been an important input into emboldening workers to organize: a sharp recession could quickly restore employer temerity to discharge workers and dampen whatever sparks in labor organizing we have now. But rising unemployment could also trigger even more militant labor activism.

One role for researchers in a moment of renewed labor activism is to build partnerships with unions new and old to study the problems of mobilization and organizing that I have highlighted in this paper, both as a laboratory for testing theories of collective action and workplace social networks and in pursuit of a subject of intrinsic policy interest. Economists have built partnerships with private companies, governments, charities, and nongovernment organizations to obtain access to administrative data and study scientific problems with randomized control trials on topics of mutual interest. Adding labor unions to this list gives us, as social scientists, a front row seat to assess which strategies of an energized labor movement might catch fire.

■ *I thank the editors, along with Daron Acemoglu, Ellora Derenoncourt, Barry Eidlin, Ethan Kaplan, Tom Kochan, Ilyana Kuziemko, Matt Mazewski, Chris Muller, Adam Reich, Ahmer Qadeer, Niha Singh, Noah Simon, Aaron Sojourner, and Eric Verhoogen for conversations and comments.*

References

Acemoglu, Daron and Jörn-Steffen Pischke. 1998. "Why Do Firms Train? Theory and Evidence." *The Quarterly Journal of Economics* 113 (1): 79–119.

Andrias, Kate. 2019. "An American Approach to Social Democracy: The Forgotten Promise of the Fair Labor Standards Act." *Yale Law Journal* 128 (3): 616–709.

Ash, Elliott, Bentley MacLeod, and Suresh Naidu. 2019. "The Language of Contract: Promises and Power in Union Collective Bargaining Agreements." Unpublished.

Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou. 2006. "Who's Who in Networks. Wanted: The Key Player." *Econometrica* 74 (5): 1403–1417.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2008. "Social Capital in the Workplace: Evidence on Its Formation and Consequences." *Labour Economics* 15 (4): 724–48.

Barth, Erling, Alex Bryson, and Harald Dale-Olsen. 2020. "Union Density Effects on Productivity and Wages." *Economic Journal* 130 (631): 1898–936.

Baumol, William J. 2012. *The Cost Disease: Why Computers Get Cheaper and Health Care Doesn't.* New Haven: Yale University Press.

Biasi, Barbara, and Heather Sarsons. 2022. "Flexible Wages, Bargaining, and the Gender Gap." *Quarterly Journal of Economics* 137 (1): 215–66.

Blanc, Eric. 2022. "How Digitized Strategy Impacts Movement Outcomes: Social Media, Mobilizing, and Organizing in the 2018 Teachers' Strikes." *Politics & Society* 50 (3): 485–518.

Boudreau, Laura, Rocco Macchiavello, Virginia Minni, and Mari Tanaka. 2021. "Union Leaders: Experimental Evidence from Myanmar." Unpublished.

Bruenig, Matt. "When McDonalds Came to Denmark." *Matt Bruening Dot Com* (**blog**)**, September** 20, 2021, https://mattbruenig.com/2021/09/20/when-mcdonalds-came-to-denmark/.

Bursztyn, Leonardo, Davide Cantoni, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang. 2021. "Persistent Political Engagement: Social Interactions and the Dynamics of Protest Movements." *American Economic Review: Insights* 3 (2): 233–50.

Calmfors, Lars, and John Driffill. 1988. "Bargaining Structure, Corporatism and Macroeconomic Performance." *Economic Policy* 3 (6): 13–61.

Clemens, Jeffrey, and Michael R. Strain. 2020. "Public Policy and Participation in Political Interest Groups: An Analysis of Minimum Wages, Labor Unions, and Effective Advocacy." Unpublished.

Cramton, Peter, and Joseph Tracy. 1998. "The Use of Replacement Workers in Union Contract Negotiations: The US Experience, 1980–1989." *Journal of Labor Economics* 16 (4): 667–701.

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

DiNardo, John, and David S. Lee. 2004. "Economic Impacts of New Unionization on Private Sector Employers: 1984–2001." *Quarterly Journal of Economics* 119 (4): 1383–441.

Dinlersoz, Emin, Jeremy Greenwood, and Henry Hyatt. 2017. "What Businesses Attract Unions? Unionization over the Life Cycle of US Establishments." *ILR Review* 70 (3): 733–66.

Dodini, Samuel, Kjell G. Salvanes, and Alexander Willén. 2021. "The Dynamics of Power in Labor Markets: Monopolistic Unions versus Monopsonistic Employers." Unpublished.

Downey, Mitch. Forthcoming. "Congressional Elections and Union Officer Prosecutions." *Review of Economics and Statistics***.**

Dube, Arindrajit, Suresh Naidu, and Adam D. Reich. 2021. "Power and Dignity in the Low-Wage Labor Market: Theory and Evidence from Walmart Workers." Unpublished.

Fantasia, Rick. 1989. *Cultures of Solidarity.* Oakland: University of California Press.

Farber, Henry S., and Bruce Western. 2001. "Accounting for the Decline of Unions in the Private Sector, 1973–1998." *Journal of Labor Research* 22 (3): 459–85.

Farber, Henry S., Daniel Herbst, Ilyana Kuziemko, and Suresh Naidu. 2021. "Unions and Inequality over the Twentieth Century: New Evidence from Survey Data." *Quarterly Journal of Economics* 136 (3): 1325–85.

Feigenbaum, James, Alexander Hertel-Fernandez, and Vanessa Williamson. 2018. "From the Bargaining Table to the Ballot Box: Political Effects of Right to Work Laws." NBER Working Paper 24259.

Ferguson, John-Paul. 2016. "Racial Diversity and Union Organizing in the United States, 1999–2008." *ILR Review* 69 (1): 53–83.

Fine, Janice. 2017. "Enforcing Labor Standards in Partnership with Civil Society: Can Co-Enforcement Succeed Where the State Alone Has Failed?" *Politics & Society* 45 (3): 359–88.

Frandsen, Brigham R. 2017. "Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable Is Discrete." In *Regression Discontinuity Designs: Theory and Applications*, Vol. 38, *Advances in Econometrics*, edited by Matias D. Cattaneo and Juan Carlos Escanciano, 281–315. Bingley: Emerald Publishing Limited.

Galeotti, Andrea, Benjamin Golub, and Sanjeev Goyal. 2020. "Targeting Interventions in Networks." *Econometrica* 88 (6): 2445–71.

Gethin, Amory, Clara Martínez-Toledano, and Thomas Piketty. 2022. "Brahmin Left versus Merchant Right: Changing Political Cleavages in 21 Western Democracies, 1948–2020." *Quarterly Journal of*

Economics 137 (1): 1–48.
González, Felipe. 2020. "Collective Action in Networks: Evidence from the Chilean Student Movement." *Journal of Public Economics* 188 (C): 1–13.
Goolsbee, Austan, and Chad Syverson. 2019. "Monopsony Power in Higher Education: A Tale of Two Tracks." NBER Working Paper 26070.
Gould, Roger V. 1995. *Insurgent Identities: Class, Community, and Protest in Paris from 1848 to the Commune.* University of Chicago Press.
Grout, Paul A. 1984. "Investment and Wages in the Absence of Binding Contracts: A Nash Bargaining Approach." *Econometrica* 52 (2): 449–60.
Hertel-Fernandez, Alexander. 2018. *Politics at Work: How Companies Turn Their Workers into Lobbyists.* Oxford: Oxford University Press.
Hertel-Fernandez, Alexander, Suresh Naidu, Adam Reich, and Patrick Youngblood. 2020. *Understanding the COVID-19 Workplace: Evidence from a Survey of Essential Workers.* New York: Roosevelt Institute.
Hertel-Fernandez, Alexander, Suresh Naidu, and Adam Reich. 2021. "Schooled by Strikes? The Effects of Large-Scale Labor Unrest on Mass Attitudes toward the Labor Movement." *Perspectives on Politics* 19 (1): 73–91.
Hertel-Fernandez, Alexander, William Kimball, and Thomas Kochan. 2022. "What Forms of Representation Do American Workers Want? Implications for Theory, Policy, and Practice." *ILR Review* 75 (2): 267–94.
Hirsch, Barry T. 2008. "Sluggish Institutions in a Dynamic World: Can Unions and Industrial Competition Coexist?" *Journal of Economic Perspectives* 22 (1): 153–76.
Hsieh, Chang-Tai, Erik Hurst, Charles I. Jones, and Peter J. Klenow. 2019. "The Allocation of Talent and US Economic Growth." *Econometrica* 87 (5): 1439–74.
Jacobs, Ken, Rebecca Smith, and Justin McBride. 2021. "State and Local Policies and Sectoral Labor Standards: From Individual Rights to Collective Power." *ILR Review* 74 (5): 1132–54.
Jin, Li, Scott Duke Kominers, and Lila Shroff. 2021. "A Labor Movement for the Platform Economy." *Harvard Business Review*, September 24. https://hbr.org/2021/09/a-labor-movement-for-the-platform-economy.
Kaplan, Greg, and Sam Schulhofer-Wohl. 2018. "The Changing (Dis-)utility of Work." *Journal of Economic Perspectives,* 32 (3): 239-58.
Kline, Patrick, Neviana Petkova, Heidi Williams, and Owen Zidar. 2019. "Who Profits from Patents? Rent-Sharing at Innovative Firms." *Quarterly Journal of Economics* 134 (3): 1343–404.
Kochan, Thomas, Janice R. Fine, Kate Bronfenbrenner, Suresh Naidu, Jacob Barnes, Yaminette Diaz-Linhart, Johnnie Kallas, et al. 2022. *U.S. Workers' Organizing Efforts and Collective Actions: A Review of the Current Landscape.* Cambridge: Worker Empowerment Research Network
Kornblum, William. 1974. *Blue Collar Community.* Chicago: University of Chicago Press.
Krackhardt, David. 1990. "Assessing the Political Landscape: Structure, Cognition, and Power in Organizations. *Administrative Science Quarterly* 35 (2): 342–369.
Kremer, Michael and Benjamin A Olken. 2009. "A Biological Model of Unions." *American Economic Journal: Applied Economics* 1 (2): 150–75.
Kuziemko, Ilyana, Nicolas Longuet Marx, and Suresh Naidu. 2022. "'Compensate the Losers?' Economy-Policy Preferences and Partisan Realignment in the US." Unpublished.
Lawrence, Lizzy, and Anna Kramer. 2021. 'How Slack and Discord Became Tools for Worker Revolt." *Protocol,* October 13. https://www.protocol.com/workplace/slack-discord-worker-protest-tools.
Logan, John. 2006. "The Union Avoidance Industry in the United States." *British Journal of Industrial Relations* 44 (4): 651–75.
Madland, David. 2021. *Re-Union: How Bold Labor Reforms Can Repair, Revitalize, and Reunite the United States.* Ithaca: Cornell University Press.
Manning, Alan. 2013. *Monopsony in Motion: Imperfect Competition in Labor Markets.* Princeton: Princeton University Press.
Marglin, Stephen A. 1974. "What Do Bosses Do? The Origins and Functions of Hierarchy in Capitalist Production." *Review of Radical Political Economics* 6 (2): 60–112.
Marx, Matt. 2011. "The Firm Strikes Back: Non-compete Agreements and the Mobility of Technical Professionals." *American Sociological Review.* 76 (5): 695–712.
Massenkoff, Maxim, and Nathan Wilmers. 2022. "Economic Outcomes of Strikers in an Era of Weak Unions. Journal of Labor Economics." Forthcoming.
McAlevey, Jane. 2016. *No Shortcuts: Organizing for Power in the New Gilded Age.* New York: Oxford University

Press.

McCarthy, Justin. 2022. "U.S. Approval of Labor Unions at Highest Point Since 1965." *Gallup*, August 30. https://news.gallup.com/poll/398303/approval-labor-unions-highest-point-1965.aspx.

Milkman, Ruth. 2006. *L.A. Story: Immigrant Workers and the Future of the U.S. Labor Movement.* New York: Russell Sage Foundation.

Naidu, Suresh. 2022. "Replication data for: Is There Any Future for a US Labor Movement?" American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E180101V1.

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

Naidu, Suresh, and Adam Reich. 2018. "Collective Action and Customer Service in Retail." *ILR Review* 71 (4): 986–1001.

Naidu, Suresh, and Aaron Sojourner. 2020. *Employer Power and Employee Skills: Understanding Workforce Training Programs in the Context of Labor Market Power.* New York: Roosevelt Institute.

Nolan, Hamilton. 2020. "A Bunch of Union Organizers Explain What's Wrong with Unions," *In These Times*, October 7. https://inthesetimes.com/article/ union-density-workers-organizing-staff-experts-public-enthusiasm.

Offe, Claus, and Helmut Wiesenthal. 1980. "Two Logics of Collective Action: Theoretical Notes on Social Class and Organizational Form." In *Political Power and Social Theory*, Vol. 1, edited by John Keane, 67–115. Cambridge: MIT Press.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Cambridge: Harvard University Press.

Philippon, Thomas. 2019. *The Great Reversal: How America Gave Up on Free Markets.* Cambridge: Harvard University Press.

Putnam, Robert D. 2000. "Bowling Alone: America's Declining Social Capital." In *Culture and Politics,* edited by Lane Crothers and Charles Lockhart, 223–34. New York: Palgrave Macmillan.

Reich, Adam, and Peter Bearman. 2018. *Working for Respect: Community and Conflict at Walmart.* New York: Columbia University Press.

Shepherd, Hana, Rebecca Roskil, Suresh Naidu, and Adam Reich. 2022. "Workplace Networks and the Dynamics of Labor Organizing." Unpublished.

Social Capital Project. 2017. *What We Do Together: The State of Associational Life in America.* Washington, DC: US Senate Joint Economic Committee.

Sojourner, Aaron J., Brigham R. Frandsen, Robert J. Town, David C. Grabowski, and Min M. Chen. 2015. "Impacts of Unionization on Quality and Productivity: Regression Discontinuity Evidence from Nursing Homes." *ILR Review* 68 (4): 771–806.

Thalen, Mikael. 2021. "Kellogg Flooded with Fake Applicants amid Company's Attempts to Replace Striking Union Workers." *Daily Dot,* December 9. https://www.dailydot.com/debug/reddit-kellogg-union-hiring-campaign/.

Uetricht, Micah, and Barry Eidlin. 2019. "U.S. Union Revitalization and the Missing 'Militant Minority.'" *Labor Studies Journal* 44 (1): 36–59.

US Court of Appeals for the District of Columbia Circuit. 2013. Conair v. NLRB.

US Supreme Court. 2014. Harris v. Quinn.

Visser, Jelle. 2019. *Trade Unions in the Balance.* Geneva: International Labour Organization Bureau of Workers' Activities. https://aias.s3.eu-central-1.amazonaws.com/website/uploads/ICTWSS_v6_1_Stata_release.dta.

Wang, Sean, and Samuel Young. 2022. "Unionization, Employer Opposition, and Establishment Closure." In *Essays on Employment and Human Capital*, PhD diss. MIT.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 29–52

Facts and Fantasies about Wage Setting and Collective Bargaining

Manudeep Bhuller, Karl Ove Moene, Magne Mogstad, and Ola L. Vestad

In most OECD countries, employers negotiate wages with labor unions. In contrast, economics textbooks fantasize about decentralized wage setting, in which individual firms and workers determine wages. In this article, we document and discuss salient features of collective bargaining systems in the OECD countries, with the goal of debunking some misconceptions and myths and revitalizing the general interest in wage setting and collective bargaining.

One myth is that collective bargaining is a single unique way of wage determination. It is not. As we shall see, there are essential differences in collective wage bargaining systems among advanced countries. Countries with comparable levels of GDP per capita, competing on the same international markets, can be very different in terms of their bargaining systems and wage structures. Even economies with the same share of unionized workers ("union density") or with the same share of workers whose terms of employment are covered by a collective agreement ("bargaining coverage") can negotiate their wages rather differently.

■ *Manudeep Bhuller is Associate Professor of Economics, University of Oslo, Oslo, Norway. He is also a Researcher, Statistics Norway, Oslo, Norway, and Research Affiliate, Centre for Economic Policy Research, London, United Kingdom. Karl Ove Moene is Professor of Economics, University of Oslo, Oslo, Norway. Magne Mogstad is Professor of Economics, University of Chicago, Chicago, Illinois. He is also a Senior Research Fellow, Statistics Norway, Oslo, Norway; Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts; and International Research Fellow, Institute for Fiscal Studies, London, United Kingdom. Ola L. Vestad is Senior Researcher, Statistics Norway, Oslo, Norway. Their email addresses are manudeep.bhuller@econ.uio.no, k.o.moene@econ.uio.no, magne.mogstad@gmail.com, and olv@ssb.no.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.29.

Major differences stem from how unions coordinate with each other. Countries like Germany, Sweden, and Norway typically have export-led pattern bargaining, in which unions in the metalworking sector and the chemical sector set the path for wage increases in private and public services. Other countries such as Israel, France, and Portugal have much less coordination across types of workers. Such differences in so-called horizontal coordination are important for how centralized the wage setting is and for how centralization works. Equally important is the level of so-called vertical coordination, reflecting whether wage bargaining takes place at level of the firm, the industry, or the nation. As we shall see, there is also a wide variation across otherwise comparable countries in terms of vertical coordination.

In general, unions tend to seek higher wages for their members, recognizing the tradeoffs in terms of possible job losses or higher consumer prices. The costs and benefits of internalizing such side effects depend critically on whether labor and output are complements or substitutes in production and demand. We argue for the plausibility of a simple but forceful principle: coordinating substitutes induces militancy, while coordinating complements induces acquiescence. We use this "Hawk-Dove" divide in union behavior to illustrate some likely implications of alternative structures of collective bargaining for wages, employment, investments, and work incentives.

Since the 1980s, most developed countries have experienced similar trends of decentralization in wage setting. Despite wide initial differences, most countries now have lower union density and less horizontal and vertical coordination than 40 years ago. Does this decentralization make theories of individually set wages more relevant? Not necessarily. It is a recent misconception that the outcome of decentralized but still collective bargaining resembles the case of individual wage determination.

Early students of labor relations, such as Beatrice and Sidney Webb (1897, p. 173), saw the difference clearly: "[T]he individual workman, applying for a job" is in a completely different position than "a group of workmen" that "sends representatives to conduct the bargaining on behalf of the whole body." Their extensive discussion in *Industrial Democracy* can perhaps be summarized by a simple rule: when each worker operates alone, local conditions of the enterprise have little impact on the wage of that person—individual characteristics of the worker are decisive. When the work group bargains in concert, however, characteristics of each worker have little impact on individual wages; instead, local conditions of the enterprise are decisive.

What we will call "Webb's rule" may also be relevant today. With collective bargaining at the firm level, equally strong unions may obtain different wages, depending on the profitability of their employer and the trade-offs they face between higher pay and lower employment. In addition, strong unions may exist in some corporations and in some plants, but not in others; some employers may have strong monopsony power, while others may have none. Such differences may lead to unequal pay for equal work and a misallocation of labor across firms and sectors, in contrast to what textbooks would predict. This assessment is important for understanding the implications of further centralization.

Are unions nothing but trouble? Clearly not. But with so many varieties of collective bargaining across countries and over time, there are some truths in both negative and positive assessments. Yet, the choice is not between uncritical blessings and overall condemnations. To insist that centralization of wage setting is generally bad for economic performance, that unions undermine important incentives, erode individualism, and demand more and more from capitalists until there is no capitalism left, misses important nuances. First, rather than excessive wage demands, wage restraint seems to be a salient feature of centralized wage setting. Second, two-tier bargaining—centrally set wages supplemented with local adjustments—can to some extent balance concerns for local incentives and flexibility in the wage setting.

Unfortunately, there is limited credible empirical evidence on the impacts of the centralization of the wage setting, and, more broadly, of the economic implications of alternative structures of collective bargaining. Indeed, much of what we know about the causes and consequences of different types of collective bargaining systems comes from theory and cross-country comparisons, subject to the usual criticism of omitted variables and endogeneity issues. For this article, instead of performing yet another cross-country comparison, we therefore analyze in the last section the wage setting in a particular country, Norway. Like many other European countries, the Norwegian collective bargaining system is based on a two-tier structure, with sectoral bargaining of wage floors or base wages followed by local bargaining at the firm level. By linking individual workers and firms to the relevant sectoral agreements, we can analyze this two-tier bargaining structure both theoretically and empirically with new register data. This analysis is centered on the question of how sectoral and local wage bargaining can be combined to trade off internalization of externalities in the wage setting with flexibility and incentives at the firm level.

Wage Setting Practices

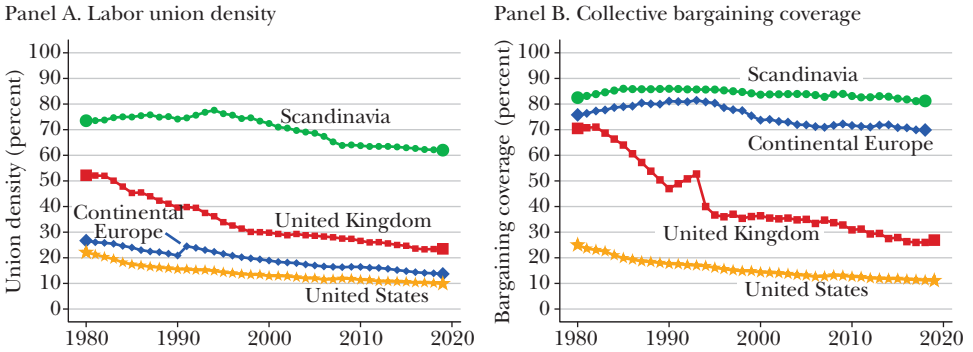
A taxonomy of wage setting practices across countries can be organized around two important dimensions: i) the level of union density and bargaining coverage and ii) the extent of vertical and horizontal coordination.

Union Density and Bargaining Coverage

A fully decentralized and individualized process of wage setting, where individual firms and individual workers determine wages, is widespread in theory (either in the form of wage posting or single-worker firm bargaining) but rare in practice. Figure 1 presents the share of workers in an economy who are union members (panel A) and the share covered by collective bargaining agreements (panel B). We present trends over time in these measures for the United States, the United Kingdom, and for different regions of Europe.

Unionization varies widely across advanced economies, as shown in panel A of Figure 1, with the highest density rates (in the Scandinavian countries) reaching

Figure 1
Trends in Union Density and Bargaining Coverage in Europe and the United States



Source: The figure is based on the OECD/AIAS database on Institutional Characteristics of Trade Unions, Wage Setting, State Intervention and Social Pacts (ICTWSS), as documented in OECD and AIAS (2021) and the OECD Labor Force Statistics (OECD 2022).
Note: This figure shows the fraction of union members (left panel) and the fraction of workers covered by collective bargaining agreements (right panel) between 1980 and 2020 for the United States and selected European countries. "Continental Europe" includes France, Germany, Spain, and Portugal, and "Scandinavian Countries" includes Norway, Sweden, and Denmark.

several times the lowest density rates (in the United States). These differences expand from 1980 to 2018, as the share of US and UK workers that are union members has steadily declined over time. Indeed, Farber et al. (2021) show that the decline in US union density started in the 1950s. As Figure 1 shows, more than half of the UK workforce was unionized in 1980, while about one-fourth of American workers were members of a union. By 2018, the union density is below 20 percent in the United Kingdom and about 10 percent in the United States.

The United States and the United Kingdom have the lowest degree of union influence. Still, in the United Kingdom, more than 10 percent of the workforce are members of one of the two largest unions: *Unite*, which organizes workers in construction, manufacturing, and transport, and *Unison*, which organizes public service workers. A decline in union membership is also found in the four Continental European countries of France, Germany, Spain, and Portugal, and, since 1990, in the Scandinavian countries of Norway, Sweden, and Denmark.

The decline in union density may seem to suggest that advanced economies have become increasingly decentralized in wage setting. However, such a conclusion would ignore that the share of workers covered by the terms of collective bargaining agreements may greatly exceed union membership. This distinction matters little in the United States. In contrast, in many Continental European countries and, to some extent in the Scandinavian countries, the share of workers covered by collective bargaining (including nonunion jobs, firms, and sectors) can substantially exceed union membership. This distinction is rooted in statutes and practices for the extension of collective bargaining agreements to workers or employers who are

not themselves members of unions or employer federations (for a detailed discussion, see Flanagan 1999). The result of these extensions is that collective bargaining agreements can directly influence the wage setting for a larger share of the workforce in these countries than the estimates of union density suggest.

Bargaining coverage in the Continental European countries has remained above 70 percent over the past four decades, despite a substantial decline in union density, as illustrated in panel B of Figure 1. In the Scandinavian countries, there is no indication of a decline in collective bargaining coverage, despite the decline in unionization since the early 1990s.

The large and increasing gap between union density and bargaining coverage in European countries is an important development that has received relatively little attention. It could be important for several reasons. For example, it might encourage nonunion workers to take a free ride on the collective bargaining efforts of union workers and thus reduce membership. If membership remains unaffected, however, extending the coverage of the union contract would raise the wage setting power of unions.

Union Coordination

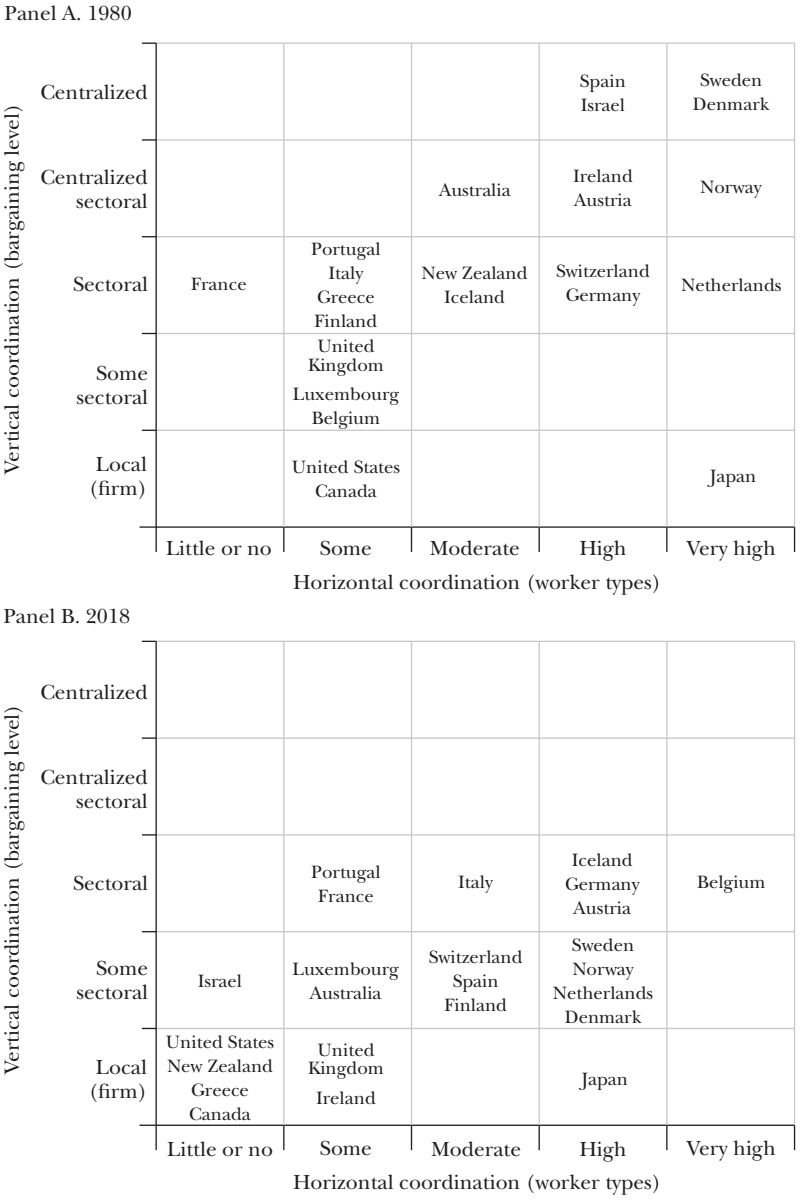
As discussed above, although a partial increase in either horizontal or vertical coordination represents more centralization, horizontal and vertical coordination capture different features of the centralization of collective bargaining. An example of unions that are coordinated horizontally, but not necessarily vertically, is the traditional craft union that organizes workers of the same craft, such as carpenters, typesetters, or shoemakers, who may well work in different firms. More coordination across unions of different crafts represents a higher level of horizontal coordination, and hence more centralization. In contrast, company unions organize many if not all types of workers within a given firm in the same union, as is common in the big corporations in Japan. The presence of company unions may imply a high level of horizontal coordination, even without coordination across unions.

Figure 2 characterizes the horizontal and vertical coordination in the collective bargaining systems of 24 advanced economies, again for 1980 (panel A) and 2018 (panel B). This characterization is based on institutional data from the OECD and the Amsterdam Institute for Advanced Labour Studies (AIAS), as documented in OECD and AIAS (2021).[1]

Most countries tend to be located along the diagonal in Figure 2. For instance, in 1980, the Scandinavian countries were at the one extreme with high degrees of

[1] For in-depth discussions of how related indexes of centralization and coordination in the collective bargaining systems across countries were originally developed in the 1970s and 1980s, see Moene, Wallerstein, and Hoel (1993) and Calmfors and Driffill (1988). Since 1994, the OECD has carried out more systematic cross-country overviews of collective bargaining systems: for a recent example, see OECD (2019a). In a widely cited cross-country overview of labor market institutions, Nickell and Layard (1999) also relied on the characterizations provided by OECD (1994) and Calmfors and Driffill (1988). We use data from the recent OECD/AIAS database because it facilitates comparisons over time and is both publicly available and well-documented (OECD and AIAS 2021).

Figure 2
Overview of Wage Setting Systems in Selected OECD Economies



Source: The figure is based on the OECD/AIAS database on Institutional Characteristics of Trade Unions, Wage Setting, State Intervention and Social Pacts (ICTWSS); see OECD and AIAS (2021).
Note: This figure provides an overview of wage setting systems in selected European countries, the United States, Canada, Australia, New Zealand, Israel, and Japan as measured in 1980 (panel A) and 2018 (panel B). Each country is categorized according to the extent of vertical and horizontal coordination in wage setting. The prevalence of vertical coordination is shown along the *y*-axis, ranging between predominantly local bargaining (at the firm level), different degrees of sectoral bargaining, and predominantly centralized bargaining. The degree of horizontal coordination is shown along the *x*-axis, ranging between little or no, some, moderate, high, and very high coordination. Each country is characterized by its predominant wage setting classification in the relevant year.

both horizontal and vertical coordination. At another extreme, the United States and Canada had little coordination, especially horizontally.

Comparing wage setting systems in 1980 and 2018, we see a clear decentralization of collective wage bargaining, with less coordination of either type. Notable examples are the Scandinavian countries that went from very high to moderate levels of vertical coordination.[2] Other examples are Greece and New Zealand, both of which shifted from moderate to low levels of both vertical and horizontal coordination.

Taken together, Figures 1 and 2 suggest that the fundamental change in how wages are set over the past few decades is a decentralization of collective wage bargaining, not a shift away from collective to individual wage setting. Below, we discuss causes, consequences, and controversies concerning this decentralization of collective wage bargaining. We also consider an important nuance ignored in Figure 2: even countries with a highly centralized bargaining system can have an important decentralized component in their wage setting. For example, collective bargaining in Scandinavia is not as centralized as the figure suggests, because the high level of coordination is combined with supplementary local wage bargaining within a two-tier framework. The Scandinavian countries are not unique in this regard. Many other developed countries, including Austria, Belgium, Italy, the Netherlands, and, more recently, Portugal and Spain (Boeri 2015), also have some version of a two-tier bargaining system.

Employer Associations and Government Involvement

Historically, unions have been considered a countervailing power against tacit collusion of employers. "We rarely hear," wrote Adam Smith (1776, I:VIII, p. 75), "of the combinations of masters, though frequently of those of workmen. But whoever imagines, upon this account, that masters rarely combine, is as ignorant of the world as of the subject." In economic analysis of unionism, however, it is too often assumed that individual firms bargain against unions with monopoly power. In practice, both employer associations and the government often play important roles in determining both the structure and the outcomes of the collective wage bargaining.

A possible reason for the one-sided focus on unions in the literature is that data on the employer side is scarce. Official statistical agencies rarely survey employers about their participation in collective bargaining, their membership in employer associations, or the extent to which pay and other employment practices are determined by collective bargaining negotiations in which they do not participate (Flanagan 1999). Of course, this lack of information does not mean that employers' organizations are irrelevant.[3]

Government may also play an important role in the bargaining between unions and employer associations. In some cases, the government's role may be

[2] See Dahl, le Maire, and Munch (2013) for a discussion of the decentralization of Danish collective bargaining system that happened in the early 1990s.

[3] For an in-depth discussion of the role of employers, see Swenson (1989).

relatively passive: it can include the provision of economic forecasts to bargaining parties, recommendations of wage-setting guidelines or norms, and appointments of mediators facilitating legal discussions and conflict resolution. In other settings, the government can play a more active role by setting minimum wages, extending collective agreements, imposing national wage schedules, imposing peace clauses on supplementary local bargaining, or ordering conflict resolutions through compulsory arbitrations.

Implications of the Structure of Collective Bargaining

The literature on collective bargaining covers a range of theoretical and empirical issues.[4] Our discussion in this section is selective and incomplete, centered around coordination and externalities in the wage setting. While this discussion will be verbal, it draws heavily on the formal results and models discussed in existing work such as Moene, Wallerstein, and Hoel (1993).

What Do Unions Care About?

There is controversy over what unions maximize. Most unions are democratic, with union members voting to influence the policies and behaviors of their organization. Theories about democratic voting have demonstrated that outcomes of elections rarely are equivalent to the maximization of some aggregate objective function, especially when heterogeneous voters are facing choices along more than one dimension. But while union members care about many issues, they are likely to have strong common interests on the topics of wages and jobs. For a private firm, economists are often willing to start with an objective function of maximizing total profits, given the belief that shareholders in big corporations are likely to have strong common interests on this subject, even though they might disagree on other subjects. Likewise, a union is typically assumed to maximize some variant of the objective function $u = u(w, L)$, with real wages w and quantity of labor L, subject to some reasonable constraint such as non-negative profits.[5]

We focus on union wage aspirations—that is, the preferred wage levels chosen by union leaders who then (at least tacitly) accept employers' right to manage

[4] For an extensive review of the literature on unions and collective bargaining, OECD (2019b) offers a useful starting point and cites many of the earlier studies since the 1980s. Freeman and Medoff (1984) is a classic work in this area. See also Elster (1989).

[5] As a concrete algebraic example, say that the unions maximize $u = (w - r)^{\gamma} L^{\beta}$, where w denotes the wage and L the employment level, γ and β are positive constants and r is the outside option wage of the members of the given union. If $\gamma = \beta = 1$ and $r = 0$, the union can be considered a bureaucratic budget maximizer, maximizing the total wage bill as a foundation to extract rents to the leadership. The case with $\beta = 1$ and $r = 0$ can be interpreted as a utilitarian union with $u = w^{\gamma} L$, maximizing the sum of union members' utility w^{γ}. In the unlikely special case with $\beta = 0$, the union maximizes union rents $(w - r)$ with no consideration of employment level L (again, subject to some reasonable constraint such as non-negative profits).

employment levels after the wage is set. Given these preferences, we consider a variety of the vertical and horizontal coordination that exists in various OECD countries.

Substitutes and Complements—the Hawk-Dove Divide

More than 100 years ago, when the United Mine Workers of America teamed up with the National Progressive Unions of Miners and Mine Laborers, basically every organized miner in the United States became a member of the same union organization. With all substitute workers organized under the same union leadership, the leadership could safely be more militant in their wage demands.

When the American Railway Union almost at the same time became an industrial union, organizing all the crafts that worked within the US railroad system, it expanded by organizing workers who were each other's complements. Consequently, the leadership of the union had to be more careful in its wage demands, as lower activities caused by higher wages to some workers would threaten the employment and wages of many other members of the same industrial union.

These two examples illustrate a simple and forceful principle, what we call the Hawk-Dove divide: Coordinating substitutes induces militancy, coordinating complements induces acquiescence.

This Hawk-Dove divide can arise for two distinct reasons. First, it can arise from workers being substitutes or complements in production, as illustrated by the miners' union versus the railroad union. More generally, consider a group of workers in a specific firm—say, steelworkers. Other metal-workers in similar firms are often substitutes to these steelworkers. Construction workers supplying inputs to metal production tend to be complements, as do shipbuilding workers who use metals as inputs. Second, the divide can arise when two groups of workers produce final outputs which are either substitutes or complements in demand. Firms within the same industry are again likely to produce outputs that tend to be substitutes in demand, implying that an increase in production by any of the firms will reduce the output price and employment for each of them. Firms in different industries, in contrast, produce outputs that are more likely to be complements in demand.

In collective wage bargaining, unions are likely to be aware of whether labor and output are complements or substitutes, and incorporate this into their wage policies and industrial actions.[6] The Hawk-Dove divide has implications also for the likelihood of conflicts between employers and unions, as the willingness to be aggressive is affected by whether unions incorporate the interests of their substitutes or their complements. The frequency of industrial actions, as measured by working days lost in wage conflict relative to the workforce, should therefore be highest when the union association primarily organizes substitutes.

[6]Horn and Wolinsky (1988) provide an insightful discussion of how the pattern of unionization depends on worker substitutability.

Meanwhile, high levels of centralization may likely lead to low levels of industrial action, since centralization beyond a certain level (vertically or horizontally) involves coordination across complements. A negative association between conflicts and centralization is consistent with data both across countries and over time, as was first demonstrated by Hibbs (1978). For example, between World Wars I and II, Sweden and Norway had little coordination in the wage setting and record-high numbers of strikes and lockouts. After World War II, however, wage setting became increasingly centralized, extending cooperation to complementary workers and sectors in both countries, and there were remarkably few strikes and lockouts in accordance with a more acquiescent union attitude.

The insight that coordination can produce acquiescence can also be seen from the impact of wage increases on prices. At the industry level, a higher wage raises the relevant producer price more than the consumer price index (which by construction reflects all prices). This imbalance can induce aggressive wage aspirations at the industry level since the costs of job losses becomes lower from every improvement in the consumer real wage. Further coordination, however, leads to acquiescence, since the impacts on producer and consumer prices become more in line as the agreements incorporate more unions and sectors.

The Salience of Union Wage Restraints

As argued above, wage restraint can be an important outcome of comprehensive coordination of wage aspirations. Increasing cooperation by incorporating different types of workers or different types of industries motivates wage moderation, to prevent either too high price effects or direct job losses among members in collaborating unions. Only when coordination shifts from the firm-level to the industry-level—when unions demand a common wage for workers who are each other's substitutes—does centralization imply militancy, with higher wages and lower employment. This observation is often missed in the discussion of collective bargaining, where it is frequently claimed that more centralized union power necessarily leads to higher wages and lower employment (for example, Baird 1984; Lindbeck and Snower 1989).

When price externalities dominate and unions coordinate vertically, both completely decentralized and centralized systems of wage bargaining can give a similar level of union wage aspirations with price-taking firms. The trade-off between real pay and jobs becomes similar in the two cases. When wages are set at the industry level, in comparison, a wage rise is less costly to the union, and it has reason to aspire for higher nominal wages. Hence, centralization can affect real wages and employment in a non-monotonic manner.

A similar result from vertical coordination also applies when employment externalities dominate. At the industry level, coordination leads to aggressive wage demands as all substitutes receive the same wage. Further coordination across branches of industries involves more coordination across complements and hence more acquiescence.

The empirical literature that aims to test whether the relationship between centralization of collective bargaining and aggregate employment or real wages is hump-shaped or monotonic has mostly relied on cross-country comparisons. In early examples, Calmfors and Driffill (1988) and Freeman (1988) found some evidence in favor of a hump-shape, while later studies have concluded differently. To date, the evidence remains mixed (for an overview of evidence, see Calmfors 2001, Table III; Moene, Wallerstein, and Hoel 1993). Key challenges in such empirical analyses are how to define a metric of centralization, how to classify each country according to this metric, and how to rule out correlated factors. For instance, Switzerland is sometimes ranked highly centralized and sometimes highly decentralized, as employers do not officially coordinate their wage offers but may do so tacitly. Japan's system of enterprise bargaining is sometimes classified as highly decentralized, while others put more weight on the coordinated wage setting across types of workers at the level of the enterprise and classify the system as centralized. The small number of observations implies that the classification of a few countries determines the overall pattern of performance—and thus whether there is a hump-shape or not.

It should also be noted that theory may predict a less pronounced hump-shape, or no hump at all, if the output market is characterized by monopolistic competition. In that case, completely centralized and decentralized wage bargaining no longer produce the same outcome. When wages are set at the firm level, a higher wage has some impact on the output price of the firm, leading to more aggressive nominal wage setting also in the decentralized case. In this context, it is also important to recall that horizontal coordination normally yields monotone wage moderation as wages are coordinated across different types of workers who are complements in either production or in demand.

Given the weak empirical and theoretical basis for a hump-shape, we argue that wage restraint is the most salient implication of the theory of coordinated wage aspirations. This insight is also in line with experiences from small open economies in Europe, where collective wage coordination is most prevalent.

Effort, Flexibility, and Investment with Central versus Local Wage Bargaining

An important cost of centralization stems from the weak flexibility and work incentives that result when wages are set independently of local performance. Conversely, an obvious advantage of decentralized collective wage bargaining is how local bargaining works as revenue-sharing that can provide powerful incentives. Local bargaining can reward local initiatives including work effort, flexibility, and skill upgrading. However, while local bargaining in the form of revenue sharing can reward current work effort and flexibility, it is likely to perform less well when it comes to the use of inputs that are sunk cost at the time of wage setting.

To illustrate these differences between local and central wage bargaining, it is useful to consider a simplified representation of the process of creative destruction. Consider therefore the stylized case where the newest technology, embodied

in the newest vintages of equipment, displaces older technologies in older vintages (for an in-depth analysis of wage coordination and creative destruction, a useful starting point is Moene and Wallerstein 1997). Once investments are made, they are sunk costs and the equipment stays in use till it becomes economically obsolete. At any point in time, there is a distribution of plants from the newest ones with the best technology to the oldest ones which just cover variable costs. As new technologies emerge, changes take place by entry of new plants and exit of old ones.

In such a setting, local bargaining means that wages are determined as a share of value added at the local plant. Even wages for homogeneous labor will thus differ across plants, with the highest wages in the most productive new plants and the lowest wages in the least productive old plants. Compared to the case with industry bargaining, which gives a uniform wage to all homogeneous workers tied to the average productivity in the industry, local bargaining works as a kind of low-wage subsidy to old inefficient plants and as a high-wage tax on the new productive plants. Therefore, firms may under-invest in new technologies and keep old equipment longer than socially optimal. Industry bargaining, in contrast, works as a tax on the least productive units and as a subsidy on the most productive. In this case, firms have incentives to invest more in modern technologies and in scrapping the old ones at an earlier stage.

Both cases may lead to a steady state with the same average growth in wages, determined by the rate of technological improvements. Yet, there are clear differences. Collective industry bargaining is expected to lead to a modernized industry with high average productivity and an egalitarian wage distribution across firms. Local bargaining should lead to a less modernized industry with a somewhat lower average productivity and with a more inegalitarian wage distribution. Thus, the bargaining system that is best for local work effort can in some respect be worst for local investments. Similarly, the flexibility entailed in local wage bargaining may work well in the case of temporary changes that require local temporary adjustments, while it may work less well with permanent shocks that require permanent adjustments. Local wage adjustments to local conditions can postpone necessary adjustments to permanent changes, delaying necessary restructuring of enterprises and industries. Advancement in one dimension can be an impediment in another. Can the two extremes be combined? This is the question of interest in the next section.

Export Led Two-Tier Bargaining in a Small Open Economy

The wage-setting practices in many small open economies in Europe, such as Norway and Sweden, are canonical examples of two-tier bargaining. In these countries, the collective bargaining system is designed to raise the competitiveness of the national economies by pursuing union wage moderation in the sectors most exposed to international competition.

Export-Led Cooperation

In Norway and Sweden, the coordinated wage-setting system came as a response to the world crisis in the 1930s. It started with a conflict over wage cuts within the union movement between sheltered and exposed unions—who were complements in both production and demand. Export-producing metalworkers stood against equally militant construction workers who to some extent were sheltered from direct foreign competition in output markets. Yet, the construction workers did produce inputs to exporting industries. During the 1930s, the metalworkers had to accept large wage cuts to stem the decline in employment. To convince construction workers to take wage cuts (to prevent high input prices for exporting industries), employers provided a helping hand: The national association of employers intervened with threats of lock-out if the construction workers did not follow the wage moderation of the metalworkers.

This was the initial step in a process of centralization of authority within the union movement in both Norway and Sweden, a process that was encouraged and supported by employers. Thus, the political coalition that prevailed in these countries after World War II—and established the so-called "centralized solidarity bargaining" system—was comprised of export-oriented workers and employers. It is unlikely that the export-oriented unions and the leadership of the union confederation would have been able to force the other high-wage unions to accept an egalitarian wage policy without the backing of employers and the threat of lockouts by employers against recalcitrant unions.

This export-led pattern-bargaining, in which unions in the export sector set the pattern for the development of wages in the rest of the economy, is controversial both in theory and in practice. The meaning of the term "pattern bargaining" has changed over time. In its early use (Webb and Webb 1897), "pattern bargaining" referred to a strategy where the most profitable industries and enterprises went first to set a pattern, to raise the wages of all workers. However, export-led pattern-bargaining is a strategy where the industries and enterprises most exposed to international competition go first to set a pattern that lowers wages, or restrains the wage increases to all workers.

There is also no consensus about just how unions of export firms can persuade unions in other firms to restrain the wage increases of their members. Is it a first-mover advantage or a repeated game argument that explains it? A form of collective rationality? Our best interpretation is simply that the role of employers remains important for maintaining the system. If some unions or industries break out from the pattern, the employers are likely to respond with threats of lock-out. Another potentially important mechanism is the role of government authorities who can take non-cooperating industries and enterprises to a "compulsory pay board" if their wage demand exceeds the export-led pattern by too much. In fact, some Norwegian unions have brought complaints to the Administrative Tribunal of the International Labour Organization (ILO) that an overuse of the "pay board," in an attempt to coordinate wage setting, violates the freedom of labor organization. Workers in the non-export industries, whose wage increases are implicitly set or constrained by the

exporting industries, often complain that they are lagging behind in the rise of real wages.

Nevertheless, export-led pattern-bargaining is frequent, and not just in the Scandinavian countries; other small open economies in Europe have also established a similar coordinated arrangement. The practice points to the possibility that unionized interests can raise the overall competitiveness of the economy.

A recent test of export-led pattern bargaining performed by Barth et al. (forthcoming) shows how union associations in countries with a high level of wage coordination have prevented local unions—who are sheltered from international competition—from reaping market power gains and raising their wages relative to workers in more exposed industries. They exploit within-country variation in exposure to trade with China in 13 European countries and find a clear pattern: In countries with wage coordination, local regions exposed to import competition from China experience no fall in employment, while in countries with uncoordinated wage setting, local regions that are exposed to import competition experience a clear fall in employment, mainly due to a reduction in manufacturing employment.

Local Supplementary Bargaining

In Scandinavian countries, the introduction of centralized systems of wage setting was later supplemented by local adjustments. The union locals wanted a say. This supplementary bargaining increased worker autonomy and the extended workplace democracy. Extreme centralization of wage setting therefore went hand-in-hand with decentralized work involvement and influence at the local level, where union leaders became substitutes for foremen and leaders at the intermediate level.

Our preferred interpretation of the details of Scandinavian two-tier bargaining is that central wage setting—the determination of the base wage q—is captured by the union wage aspirations as discussed above, while the supplementary bargaining at the local level provides wage drift d, implying that the local wage is $w = q + d$.[7]

The wage drift is best understood as a form of negotiated revenue-sharing at the level of the firm or the plant. At this local level, however, there are restrictions on the degree and type of industrial conflict. Norway and Sweden have a "peace clause" in the main agreements between the peak associations of labor and capital, which forbids strikes and lock-outs between the time when a central agreement is reached and the start of the negotiations for a new agreement. The implicit threats that can be used at the local level are therefore restricted to "work-to-rule" actions, in which workers slow production via strict observation of the letter of the rules, without reducing production by so much that the firm responds by laying off workers. This approach will plausibly yield the local unions a lower revenue share than they would have obtained with viable strike threats.

[7]The definition of the term "wage drift" varies across papers. It is sometimes defined as the difference and other times as the change in the wage actually paid to a worker as compared to her base wage. Throughout the paper, we let wage drift denote the difference (not the change) in the wage actually paid to a worker as compared to a base wage.

The restriction on the local use of industrial conflicts in a two-tier system has two major implications. First, it ties wages to local productivity, but with a lower elasticity than in the pure local bargaining case in which strikes are permitted, as the share of the revenues that the union obtains is lower. Nevertheless, a linkage from local wages to firm profits can create some incentives for good work performance and involvement at the firm level. Second, pure local bargaining runs a risk of subsidizing old and inefficient firms with lower wages, while imposing an implicit tax in the form of higher wages on firms that make productive new investments. Two-tier bargaining can therefore strike a balance between the concerns for work incentives and investment incentives.

Does the flexibility of two-tier bargaining lead to the same outcomes as decentralized collective bargaining? Particularly when the drift is high relative to total changes in wages, it might seem as though the answer is yes. But on the contrary, we argue that the two-tier system functions as a centralized system of wage setting whether the supplementary wage increases are higher than the centrally negotiated base wage increases or not. At the central level, the negotiators can foresee (or make a qualified guess on) the average wage drift that will come on top of the centrally negotiated base wage, and they can incorporate this drift in their wage aspirations. Obviously, the negotiators can only incorporate the typical or average drift, implying that workers in the most productive enterprises obtain a higher wage than what lies in the implicit bargaining goal, while workers in less efficient firms obtain less than the bargaining goal. Nevertheless, the structure and level of wages are determined by the union aspirations at the central level. Holden (1998) offers an in-depth theoretical and empirical discussion of both wage drift and downward wage rigidity under centralized bargaining in the Nordic countries.

Empirical Illustrations in the Case of Norway

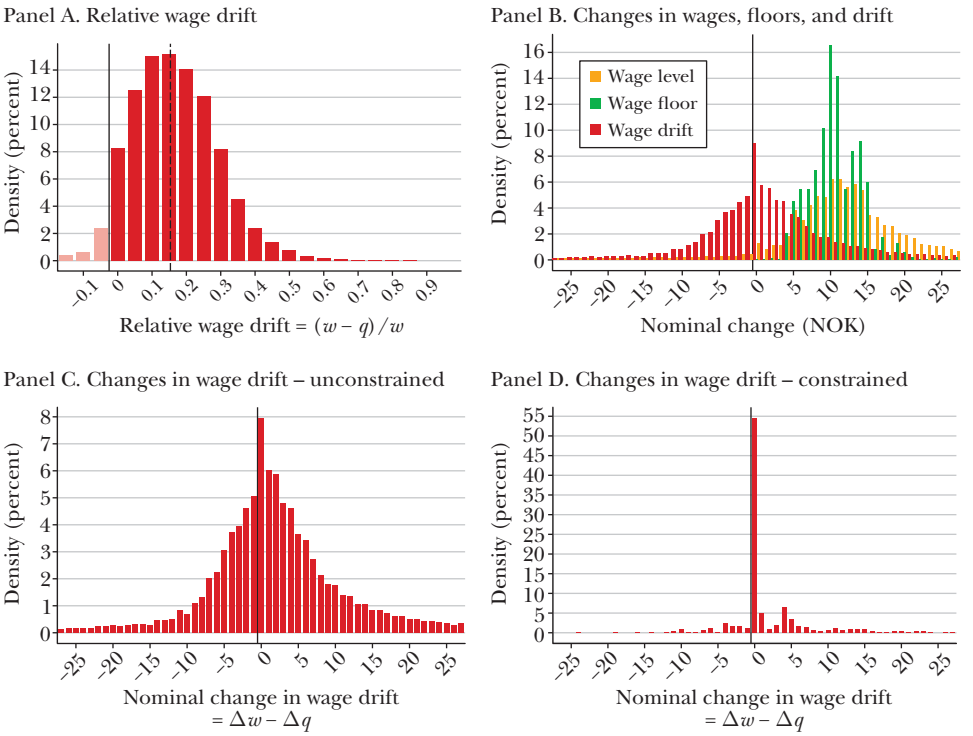
We now draw on high-quality micro data to illustrate the "anatomy" of the wage setting in Norway, a small open economy with a two-tier bargaining system.[8] We present empirical evidence on composition of wages and changes in wages, wage inequality within and between industries, and pattern bargaining. The goal is to tie the theory of collective bargaining discussed above to the wage structure we observe in an actual economy.

Wage Floors and Drift with Two-Tier Bargaining

Above we emphasized how the base wage q acts as a wage floor, and that wage drift d (equal to $w - q$) is non-negative, but not the same for all workers. Figure 3 confirms this pattern empirically: It shows distributions of relative wage drifts and

[8] Details about data sources, variables, and the procedure for linking of individual workers to job-specific wage floors are provided in the online Appendix. See also Card and Cardoso (2021), who use similar data to analyze the collective bargaining system in Portugal.

Figure 3
Distributions of Wage Drift and Changes in Wages, Floors, and Drift



Source: Our calculations based on administrative data from Statistics Norway's Population Statistics (Statistisk sentralbyrå 2000a); Education Statistics (Statistisk sentralbyrå (2000b; 2001); Wage Survey Statistics (Statistisk sentralbyrå, 2011; 2020a); and Employer-Employee Register and Employment Statistics (Statistisk sentralbyrå, 2010; 2020b). These administrative registers were linked to transcribed data on wage floors from archival records and restricted-access data on collective bargaining coverage from the Confederation of Norwegian Enterprise (NHO 2022). See further details in the online Appendix

Note: This figure shows distributions of relative wage drifts and changes in nominal wages, floors, and drifts for workers covered by a collective bargaining agreement. In panel A, wage drifts are measured for each even year in the period 2010–2018. For each of these years, the sample includes full-time workers who did not change jobs in the same year. The figure shows wage drift measured as a fraction of total wages. The changes in panels B–D are calculated based on observed changes in nominal wages, drifts, and floors between two consecutive even years for full-time workers who did not change jobs between the two years, and reported in Norwegian kroner (NOK). For panels C and D, "unconstrained workers" are defined as workers earning wages strictly above the wage floor associated with their job in year $t-2$, while the "constrained workers" are those earning wages equal to the associated wage floor in year $t-2$. Observations of hourly wages above 2,000 Norwegian kroner, below 50 kroner, or below 20 percent of the associated wage floor are excluded.

nominal changes in wages, floors, and drifts for workers covered by a collective bargaining agreement.

Panel A of Figure 3 shows estimates of relative wage drift in Norway, using data for even years from 2010 through 2018 for workers covered by a collective

bargaining agreement. The reason for looking at even years is that the main negotiations between unions and employer federations happen every even year (that is, not every year). The graph reveals that although about 10 percent of workers are paid wages equal to or close to the base wage that apply to their jobs, most workers receive substantial wage premiums above this floor. For a typical worker, this wage drift corresponds to about 15 percent of the wage.

There seems to be two types of rigidities in the Norwegian wage structure. One is about the *level* of the wage drift d, which is almost never negative (see panel A of Figure 3). The other is about the *changes* in wages. Nominal wages and wage floors are (almost) never adjusted downwards, likely reflecting downward rigidity in nominal wages.[9] Panel B of Figure 3 shows distributions of changes in nominal wages, changes in wage floors, and changes in wage drift between two consecutive even years for a sample of job stayers. Changes in wage floors are never negative, while only about 8 percent of all workers received a nominal wage cut during a two-year period. By contrast, the distribution of changes in wage drift resembles a bell shape, albeit with a clear spike at zero and a somewhat smaller left tail than right tail.

Panels C and D show the distributions of changes in wage drifts for two separate groups of job-stayers. The "unconstrained workers" in panel C are workers earning wages strictly above the wage floor associated with their job, while the "constrained workers" in panel D are those earning wages approximately equal to the associated wage floor. Less than 5 percent of workers are constrained in this manner, which reflects the importance of wage drift in our data. The distribution in panel C resembles the distribution for the full sample of job stayers in panel B, with a spike at zero showing that only 8 percent of the unconstrained workers received a nominal wage increase exactly equal to the increase in their wage floor. By contrast, more than 50 percent of the workers who earned wages equal to their wage floors in the previous period earn wages equal to their wage floors also in the present one. Most constrained workers thus also remain constrained two years ahead.

Wage Inequality with Two-Tier Bargaining

Centralized collective bargaining is likely to affect the extent of wage dispersion, both across and within industries. With Norway's strong horizontal coordination in wage setting across industries, one would expect inter-industry wage differentials for observationally similar workers to be limited. When negotiators internalize price and employment externalities across types of workers, we obtain base wages that tend toward equal pay for observationally similar workers. However, inter-industry wage differentials may persist, primarily due to systematically different quantities of wage drift in industries with different labor productivity. Indeed, a decomposition of the variance of wages between and within industries reveals that about 40 percent

[9] Nominal wage rigidities are also evident in many other countries and wage settings. For more discussion and evidence on downward nominal wage rigidity, see, for example, Dickens et al. (2007) in this journal and Grigsby, Hurst, and Yildirmaz (2021).

of the variation in wages in Norway can be attributed to differences in wages across industry-wide collective bargaining agreements. The structure of collective bargaining should also matter for intra-industry wage differentials. Norway's strong vertical coordination should imply limited dispersion in wages across firms within the same industry. However, the two-tier bargaining structure allows for local wage supplements, which could also lead to persistent differences in wages across firms within the same industry, depending on firm-specific productivity.

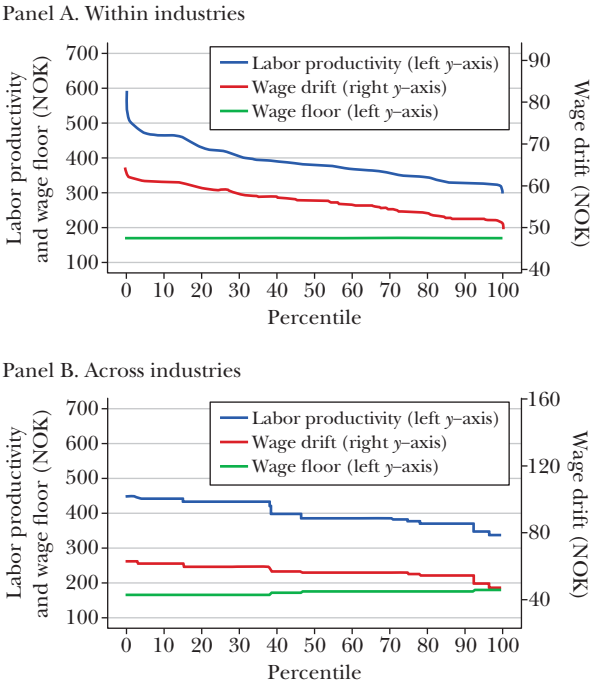
Figure 4 explores the relationship between wages and labor productivity within and across industries. We define firm average labor productivity as firm revenues minus input costs and changes in the value of stock of produced goods. Wage floors and drifts are measured net of observable worker characteristics and workers are sorted by labor productivity, with workers employed in the most productive firms to the left and workers employed in the least productive firms to the right. Panel A shows intra-industry differences in labor productivity, wage floors, and wage drift for all collective bargaining agreements covered by our sample (net of average differences across agreements). Note that both labor productivity (blue line) and wage floor (green line) are indexed in Norwegian kroner (NOK) along the left y-axis, while wage drift (red line) is shown along the right y-axis.

Figure 4 illustrates two key features of a collective wage setting system with two-tier bargaining. First, the centrally negotiated base wage establishes a common wage floor in each industry. Second, the locally negotiated wage drifts produce significant differences in wages across workers within the same industry, depending on the productivity of the firm in which they are employed. As the theory predicts, the least productive firms pay wages that are approximately equal to their labor productivity, while the most productive firms pay wages that are much lower than labor productivity, earning positive (quasi-)rents on the workers. This evidence is consistent with how two-tier bargaining can reflect a compromise between work and investment incentives, as discussed above.

Panel B of Figure 4 shows the relationship between average wages and average labor productivity *across* industries. Consistent with the theory of how wage coordination across workers who are complements (in production or in demand) leads to wage restraint, there is little evidence of a systematic relationship between the wage floors and the average productivity of the industries. If anything, moving from high- to low-productivity industries, we see a decline in average wage drifts and a slight increase in wage floors. This pattern is consistent with a wage-setting system in which the base wage is set slightly higher in industries where one expects a lower average drift. Overall, the relatively small differences in wage floors across high- and low-productivity industries can be interpreted as evidence of strong horizontal coordination across industries in Norway's collective bargaining system.

A concern with Figure 4 is that it only uses cross-sectional data, which means that the wage differentials may reflect unobserved differences in the quality of workers. Interestingly, if we instead use the panel data available to us in this setting to study the relationship between wages and changes in productivity within and

Figure 4
Labor Productivity, Wage Floors, and Wage Drift



Source: Our calculations based on administrative data from Statistics Norway's Population Statistics (Statistisk sentralbyrå 2000a); Education Statistics (Statistisk sentralbyrå (2000b; 2001); Wage Survey Statistics (Statistisk sentralbyrå, 2011; 2020a); Employer-Employee Register and Employment Statistics (Statistisk sentralbyrå, 2010; 2020b); and Firm Accounts Statistics (Statistisk sentralbyrå 2020c). These administrative registers were linked to transcribed data on wage floors from archival records and restricted-access data on collective bargaining coverage from the Confederation of Norwegian Enterprise (NHO 2022). See further details in the online Appendix.
Note: This figure shows average labor productivity, wage floors, and wage drift by percentiles in the worker-weighted distribution of labor productivity, ranked in descending order, with wage floors and drifts measured net of observable worker characteristics. The lines are estimates from kernel (local constant) regressions of labor productivity, wage floors, and wage drift, respectively, on percentile group indicators. Panel A shows firm/worker level measures (net of differences across collective bargaining agreements) for firms and workers covered by any of the 18 collective bargaining agreements in our sample. Panel B shows agreement level average labor productivity, wage floors, and wage drifts. Labor productivity, wage floors, and wage drifts are measured for each even year in the period 2010–2018 (net of differences across years), and for each of these years, the sample includes firms with at least five workers in the relevant year and positive value added in the surrounding five-year period. The sample of firms is truncated at the fifth and ninety-fifth percentile in the distribution of labor productivity. Wage floors and drifts are measured for all full-time workers between the ages of 25 and 60 who did not change jobs in the relevant year, and wages are winsorized at the 2.5th and 97.5th percentiles.

across industries, we find similar patterns. Positive productivity shocks are associated with higher wages, regardless of whether the shocks are common to all industries, specific to certain industries, or specific to certain firms within an industry. And while common productivity shocks tend to raise wages primarily through adjustments of

wage floors, industry- and firm-specific shocks are transmitted to wages in the form of changes in the locally negotiated wage drift.

The Salience of Pattern Bargaining

Our discussion about export-led coordination has highlighted how this type of pattern bargaining can allow the industries and enterprises most exposed to foreign competition to set a pattern of wage increases that applies to the rest of the economy. In the Norwegian context, export-oriented manufacturing has traditionally functioned as the "front runner" in the centralized collective bargaining system, so that the wage settlements in the manufacturing agreement set norms for wage settlements that take place in the other collective bargaining agreements (for a historical overview of this "front runner" system, see Nymoen 2017, Section 2.5).

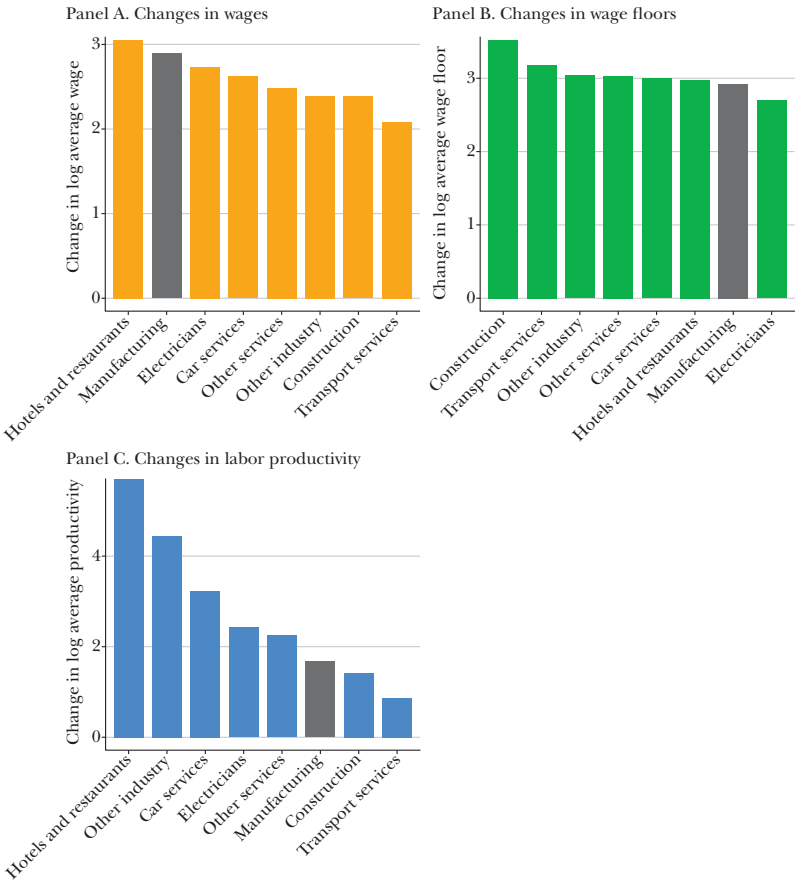
In Figure 5, we focus on eight major industries, where each industry can have multiple collective bargaining agreements. Panel A shows annual growth rates in mean wages, averaged between years 2010 and 2018, for each industry. Consistent with manufacturing being the "front runner," we find the second highest average wage growth in this industry. By comparison, panel B suggests limited differences in the growth rates of negotiated wage floors across industries. The "wage growth premium" in favor of the manufacturing industry becomes even more striking when we consider industry differences in the growth of mean labor productivity in panel C, where manufacturing has had among the lowest growth rates. Despite the low growth in manufacturing productivity, Norwegian manufacturing has been able to retain a high growth in mean wages. We interpret this as empirical support of a strong influence of export-led pattern bargaining in the Norwegian system of collective bargaining. The sustainability and economic consequences (for example, in terms of (mis)allocation of labor, aggregate productivity, wage inequality) of this system are important but largely unresolved questions.

Concluding Remarks

We have documented and discussed salient features of collective bargaining systems in the OECD countries, with the goal of debunking some misconceptions and myths and revitalizing the general interest in wage setting and collective bargaining. We hope that such an interest may help close the gap between how economists tend to model wage setting and how wages are actually set. The textbook models of competitive labor markets, monopsony, and search and matching all assume a decentralized wage setting where individual firms and workers determine wages. In most advanced economies, however, it is common that firms or employer associations bargain with unions over wages, producing collective bargaining systems. The characteristics of these systems vary substantially across advanced economies, with regards to both the scope and the structure of the collective bargaining.

Figure 5

Annual Changes in Average Wages, Wage Floors, and Labor Productivity



Source: Our calculations based on administrative data from Statistics Norway's Population Statistics (Statistisk sentralbyrå 2000a); Education Statistics (Statistisk sentralbyrå (2000b; 2001); Wage Survey Statistics (Statistisk sentralbyrå, 2011; 2020a); Employer-Employee Register and Employment Statistics (Statistisk sentralbyrå, 2010; 2020b); and Firm Accounts Statistics (Statistisk sentralbyrå 2020c). These administrative registers were linked to transcribed data on wage floors from archival records and restricted-access data on collective bargaining coverage from the Confederation of Norwegian Enterprise (NHO 2022). See further details in the online Appendix.

Note: This figure shows annual changes in the log of collective bargaining agreement-level average wages, wage floors, and labor productivity for different groups of agreements. Average wages, wage floors, and labor productivity are measured for each even year in the period 2010–2018, and changes in the log of these averages are calculated for each pair of successive even years. The annual changes shown in the figure are obtained by dividing the two-year changes by two and multiplying by 100. For each even year in the period 2010–2018, the sample includes firms with at least five workers and positive value added in the relevant year. The sample of firms is truncated at the fifth and ninety-fifth percentile in the distribution of labor productivity. Wage floors and drifts are measured for all full-time workers between the ages of 25 and 60 who did not change jobs in the relevant year, and wages are winsorized at the 2.5th and 97.5th percentiles. The collective bargaining agreement groups are defined as follows: Manufacturing (Manufacturing, Textile and Confection, Technology and Data); Other Industry (Cartonage, Meatpacking Industry, Construction Materials Industry); Construction (Construction Trades, Private Construction Contractors); Electricians (Electricians Trade); Car Services (Car Services); Transport Services (Bus Industry, Freight Forwarding, Transport Firms); Hotels and Restaurants (The National Agreement—for hotel and restaurant workers); Other Services (Cleaning, Private Security).

Understanding the causes and consequences of different wage-setting practices and work organization has a long history in labor economics. However, these questions have, over time, become less fashionable. Instead, many labor economists have shifted attention to understanding the relative importance of individual determinants of wages given the wage-setting practice in the economy of study.

For example, in the context of the lightly unionized US economy, numerous studies have sought to identify a causal effect of the union wage premium—that is, how much more an otherwise identical American worker is paid as a result of union membership. Much effort has gone into improving the research design of such studies.[10] While these improvements have been important, the results of these kinds of quasi-experimental studies are only informative about how a marginal increase in union membership, given the wage-setting practices in the American economy, would benefit the workers entering a union. More generally, a study focused on changing an individual determinant of wages, while holding the overall system of wage setting fixed, cannot tell us about the systemic effects of broader changes in the wage setting system. We suspect that real progress in the study of wage-setting institutions broadly understood will require a shift in research towards careful modeling of the actual institutional setting and tighter connections between data and theory.

■ *We thank Arnstein Vestre, Karstein Sørlien, Fanny Berg, Maria Naomi Christensen, and Maren L. Eilertsen for excellent research assistance. Manudeep Bhuller received financial support from the Research Council of Norway through grant 275123 (UNIFRIC) and the European Research Council through grant 101043127 (LABFLEX). Magne Mogstad and Ola Vestad received financial support from the Research Council of Norway through grant 295901. We are also grateful for constructive comments from the editors and Steinar Holden who read an earlier version.*

[10] A few prominent examples of such studies include the following: Ashenfelter (1978) constructs control groups for union members based on industry, race, and worker type (like craftsmen, operatives, laborers); Freeman (1984) compares wage rates for the same individual who changes unionization status over time; Lemieux (1998) compares wage rates for the same individual who holds two jobs, one of which is unionized, and the other is not; Krashinsky (2004) compares wage rates of identical twins, one who is unionized and one who is not; and DiNardo and Lee (2004) use a regression discontinuity design that takes advantage of the fact that new unionizations often occur as a result of a secret ballot election.

References

Ashenfelter, Orley. 1978. "Union Relative Wage Effects: New Evidence and a Survey of their Implications for Wage Inflation." In *Econometric Contributions to Public Policy*, edited by Richard Stone and William Peterson, 31–63. London: Palgrave Macmillan.

Baird, Charles W. 1984. *Opportunity or Privilege: Labor Legislation in America*. Studies in Social Philosophy & Policy. Piscataway: Transaction Publishers.

Barth, Erling, Henning Finseraas, Anders Kjelsrud, and Karl Ove Moene. Forthcoming. "Hit by the Silk Road: How Wage Coordination in Europe Mitigates the China Shock." *Scandinavian Journal of Economics*.

Boeri, Tito. 2015. "Perverse Effects of Two-Tier Wage Bargaining Structures." *IZA World of Labor*. Article 101, January, 1–10.

Bhuller, Manudeep, Karl Ove Moene, Magne Mogstad, and Ola L. Vestad. 2022. "Replication data for: Facts and Fantasies about Wage Setting and Collective Bargaining." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E178882V1.

Calmfors, Lars. 2001. "Wages and Wage-Bargaining Institutions in the EMU—A Survey of the Issues." *Empirica* 28 (4): 325–52.

Calmfors, Lars, and John Driffill. 1988. "Bargaining Structure, Corporatism and Macroeconomic Performance." *Economic Policy* 3 (6): 13–61.

Card, David, and Ana Rute Cardoso. 2021. "Wage Flexibility under Sectoral Bargaining." NBER Working Papers No. 28695.

Dahl, Christian M., Daniel le Maire, and Jakob R. Munch. 2013. "Wage Dispersion and Decentralization of Wage Bargaining." *Journal of Labor Economics* 31 (3): 501–33.

Dickens, William T., Lorenz Goette, Erica L. Groshen, Steinar Holden, Julian Messina, Mark E. Schweitzer, Jarkko Turunen, and Melanie E. Ward. 2007. "How Wages Change: Micro Evidence from the International Wage Flexibility Project." *Journal of Economic Perspectives* 21 (2): 195–214.

DiNardo, John, and David S. Lee. 2004. "Economic Impacts of New Unionization on Private Sector Employers: 1984-2001." *Quarterly Journal of Economics* 119 (4): 1383–441.

Elster, Jon. 1989. *The Cement of Society: A Study of Social Order*. New York: Cambridge University Press.

Farber, Henry S., Daniel Herbst, Ilyana Kuziemko, and Suresh Naidu. 2021. "Unions and Inequality over the Twentieth Century: New Evidence from Survey Data." *Quarterly Journal of Economics* 136 (3): 1325–85.

Flanagan, Robert J. 1999. "Macroeconomic Performance and Collective Bargaining: An International Perspective." *Journal of Economic Literature* 37 (3): 1150–75.

Freeman, Richard B. 1984. "Longitudinal Analyses of the Effects of Trade Unions." *Journal of Labor Economics* 2 (1): 1–26.

Freeman, Richard B. 1988. "Labour Market Institutions and Economic Performance." *Economic Policy* 3 (6): 63–80.

Freeman, Richard B., and James L. Medoff. 1984. *What Do Unions Do?* New York: Basic Books.

Grigsby, John, Erik Hurst, and Ahu Yildirmaz. 2021. "Aggregate Nominal Wage Adjustments: New Evidence from Administrative Payroll Data." *American Economic Review* 111 (2): 428–71.

Hibbs, Douglas A., Jr. 1978. "On the Political Economy of Long-Run Trends in Strike Activity." *British Journal of Political Science*, 8 (2): 153–75.

Holden, Steinar. 1998. "Wage Drift and the Relevance of Centralised Wage Setting." *Scandinavian Journal of Economics* 100 (4): 711–31.

Horn, Henrik, and Asher Wolinsky. 1988. "Worker Substitutability and Patterns of Unionisation." *Economic Journal* 98 (391): 484–97.

Krashinsky, Harry A. 2004. "Do Marital Status and Computer Usage Really Change the Wage Structure?" *Journal of Human Resources* 39 (3): 774–91.

Lemieux, Thomas. 1998. "Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection." *Journal of Labor Economics*, 16 (2): 261–91.

Lindbeck, Assar, and Dennis J. Snower. 1989. *The Insider-Outsider Theory of Employment and Unemployment*. Cambridge, MA: MIT Press.

Moene, Karl Ove, and Michael Wallerstein. 1997. "Pay Inequality." *Journal of Labor Economics* 15 (3): 403–30.

Moene, Karl Ove, Michael Wallerstein, and Michael Hoel. 1993. "Bargaining Structure and Economic Performance." In *Trade Union Behavior, Pay Bargaining and Economic Performance*, Part II, edited by Robert Flanagan, Karl Ove Moene, and Michael Wallerstein, 62–132. Oxford: Oxford University Press.

Næringslivets Hovedorganisasjon. "Lønn og tariff." Oslo: NHO. https://www.nho.no/lonn-og-tariff/ (accessed August 29, 2022).

Nickell, Stephen, and Richard Layard. 1999. "Labor Market Institutions and Economic Performance." In *Handbook of Labor Economics,* Vol. 3, edited by Orley Ashenfelter and David Card, 3029-84. Amsterdam: Elsevier.

Nymoen, Ragnar. 2017. "Between Institutions and Global Forces: Norwegian Wage Formation since Industrialisation." *Econometrics* 5 (1): 1–54.

OECD. 1994. "Collective Bargaining: Levels and Coverage." In *Employment Outlook*, 167–208. Paris: OECD.

OECD. 2019a. "Facing the Future of Work: How to Make the Most of Collective Bargaining." In *Employment Outlook*, 189–234. Paris: OECD.

OECD. 2019b. *Negotiating Our Way Up: Collective Bargaining in a Changing World of Work*. Paris: OECD.

OECD. 2022. "Labor Force (Indicator). Paris: OECD. doi: 10.1787/ef2e7159-en (accessed August 29, 2022).

OECD, and AIAS. 2021. "OECD/AIAS ICTWSS Database: Note on Definitions, Measurement and Sources." Paris: OECD. https://www.oecd.org/employment/ictwss-database.htm (accessed September 28, 2022).

Smith, Adam. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. Chicago: University of Chicago Press, 1976.

Statistisk sentralbyrå. 2000a. "Dokumentasjon av BESYS—befolkningsstatistikksystemet." Notater 2000/24.

Statistisk sentralbyrå. 2000b. "Norsk standard for utdanningsgruppering." Norges offisielle statistikk, C 617.

Statistisk sentralbyrå. 2001. "Utdanningsstatistikk: Dokumentasjon 2000 av den individbaserte utdanningsstatistikken." Norges offisielle statistikk, C 645.

Statistisk sentralbyrå. 2010. "Registerbasert sysselsettingsstatistikk: Dokumentasjon." Notater 8/2010.

Statistisk sentralbyrå. 2011. "Lønnsstatistikk og årslønn: Dokumentasjon av beregningsopplegg for årslønn." Notater 12/2011.

Statistisk sentralbyrå. 2020a. "Lønn: Data om lønnsmottakeres månedslønn, som avtalt månedslønn, overtid, bonus og uregelmessige tillegg." https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/lonn (accessed August 29, 2022).

Statistisk sentralbyrå. 2020b. "A-ordningen: Data om personers tilknytning til arbeidsmarkedet." https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/a-ordningen (accessed August 29, 2022).

Statistisk sentralbyrå. 2020c. "Regnskap: Data fra resultat- og balanseregnskap for ikke-finansielle aksjeselskaper." https://www.ssb.no/data-til-forskning/utlan-av-data-til-forskere/variabellister/regnskap (accessed August 29, 2022).

Swenson, Peter. 1989. *Fair Shares: Unions, Pay and Politics in Sweden and West Germany*. Ithaca: Cornell University Press.

Webb, Sidney, and Beatrice Webb. 1897. Industrial Democracy. London: Longmans, Green, and Co.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 53–80

The German Model of Industrial Relations: Balancing Flexibility and Collective Action

Simon Jäger, Shakked Noy, and Benjamin Schoefer

Germany—the world's fourth-largest economy—has remained partially insulated from the growing labor market challenges faced by the United States and other high-income countries. In many advanced economies, the past few decades have seen sustained increases in earnings inequality, a fall in the labor share, the disappearance of "good jobs" in manufacturing, the rise of precarious work, and a deterioration in the power of organized labor and individual workers.[1] These developments threaten to prevent economic growth from translating into shared prosperity.

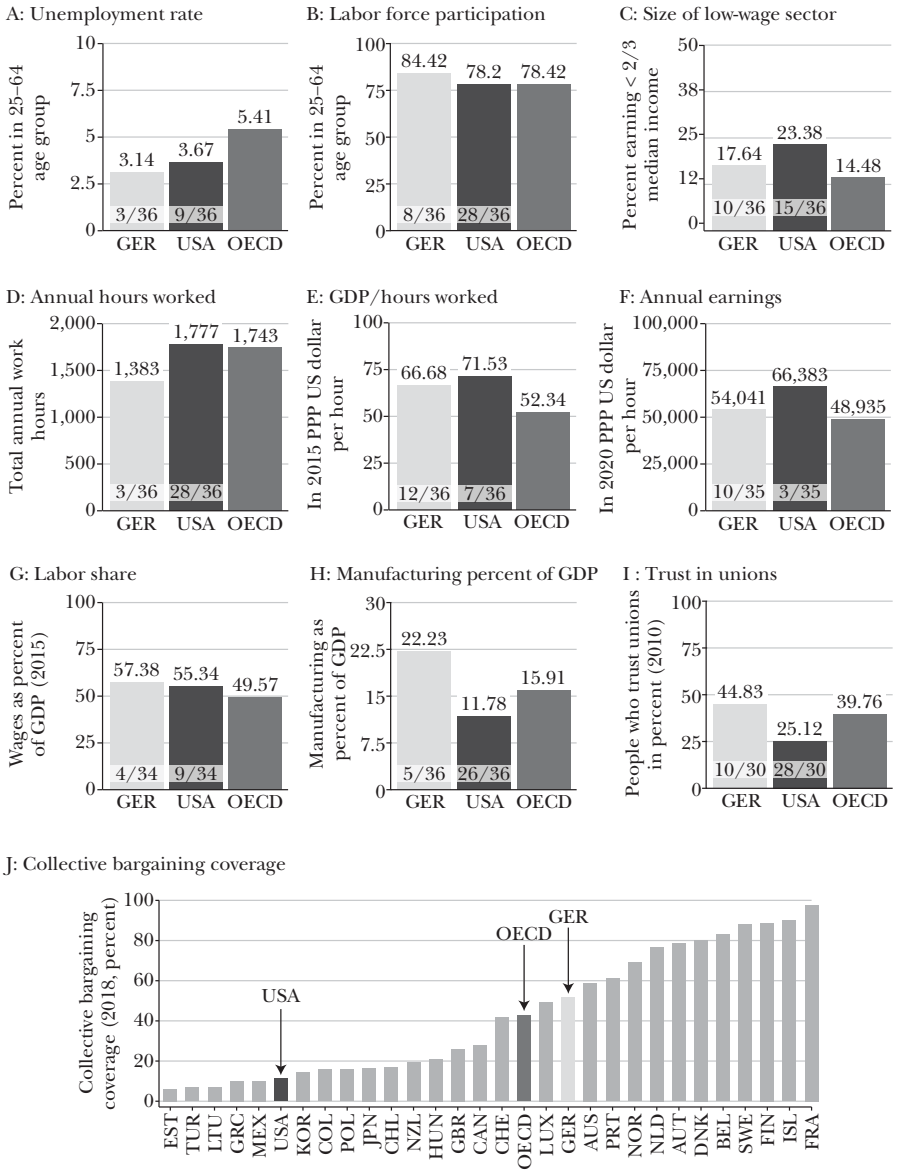
Compared to the United States, German organized labor has remained strong, as shown in Figure 1. Half of German workers are covered by a collective bargaining agreement, compared to 6.1 percent of private-sector Americans (Bureau of Labor Statistics 2022). Trust in unions is almost twice as high in Germany compared to the United States. Employees in Germany work fewer hours, the country's low-wage sector is 25 percent smaller, and labor's share of national income is higher. The German manufacturing sector still makes up almost one-quarter of GDP (compared

■ *Simon Jäger is Associate Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts, and CEO of IZA Institute of Labor Economics, Bonn, Germany. Shakked Noy is a PhD student in economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Benjamin Schoefer is Associate Professor of Economics, University of California, Berkeley, in Berkeley, California. Jäger and Schoefer are both Faculty Research Fellows, National Bureau of Economic Research, Cambridge, Massachusetts. Their emails are sjaeger@mit.edu, snoy@mit.edu, and schoefer@berkeley.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.53.

[1] See, for example, Autor (2014) and Chancel et al. (2022); Karabarbounis and Neiman (2014); Autor, Dorn, and Hanson (2013) and Acemoglu and Restrepo (2020); Weil (2014); and Stansbury and Summers (2020) and Farber et al. (2021).

Figure 1
The German Labor Market in International Comparison



Source: Unemployment = annual percentage of unemployed aged 25 to 64 (OECD 2022f). Employment rate = employed aged 25 to 64/population aged 25 to 64 (OECD 2022e). Low wage sector share = share of employees earning less than two-thirds of the annual median wage (OECD 2022h). Annual working hours = total annual working hours/employed population (OECD 2022d). Annual wage = total annual wage in constant prices 2020 US dollars at the purchasing power parity exchange rate (OECD 2022a). GDP per hour = annual GDP in constant prices 2020 US dollars at the purchasing power parity exchange rate/total annual working hours (OECD 2022c). Labor share = employee compensation as share of GDP (OECD 2022b). Manufacturing share = manufacturing sector output as share of GDP (OECD 2022g). Trust in unions = share of people who tend to trust unions (Germany) or who are "greatly" or "quite a lot" confident in unions (United States) (OECD 2019). Bargaining coverage = share of workers covered by a collective agreement (Visser 2021; OECD 2021).

Note: Unless otherwise noted, the data are for 2019. Numbers above the bars denote the heights of the bars; the numbers at the bottom of the bars denote the US/Germany's rank in the OECD in terms of each measure (with ranks closer to 1 being "better" for all measures). Variation in the total number of OECD countries is due to missing data (for example, no data for annual wages for Turkey) and different years (manufacturing share data from 2018). Both manufacturing and labor share OECD averages were calculated by the authors from OECD data.

to 12 percent in the United States). Germany has one of the highest robot penetration rates in the world (International Federation of Robotics 2017)—yet in contrast to the United States (Acemoglu and Restrepo 2020), robotization has not led to net employment declines in Germany, especially in areas with high union strength (Dauth et al. 2021). At the same time, relative to other OECD countries—many of which, like France or Italy, have maintained even higher collective bargaining coverage through more rigid bargaining systems—the German labor market features low unemployment and high labor force participation (though also a larger low-wage sector).

Motivated by these facts, observers and policymakers in other countries have paid increasing attention to the German model of industrial relations (for example, Anderson 2012; *The Economist* 2017, 2020; Matthews 2019; Strine, Kovvali, and Williams 2021). In the 2020 US Democratic primaries, the policy platforms of several candidates contained proposals explicitly based on German labor market institutions (Campbell 2019). And American workers, frustrated with their lack of voice, exhibit demand for workplace representation mechanisms in the mold of the German system (Hertel-Fernandez, Kimball, and Kochan 2022).

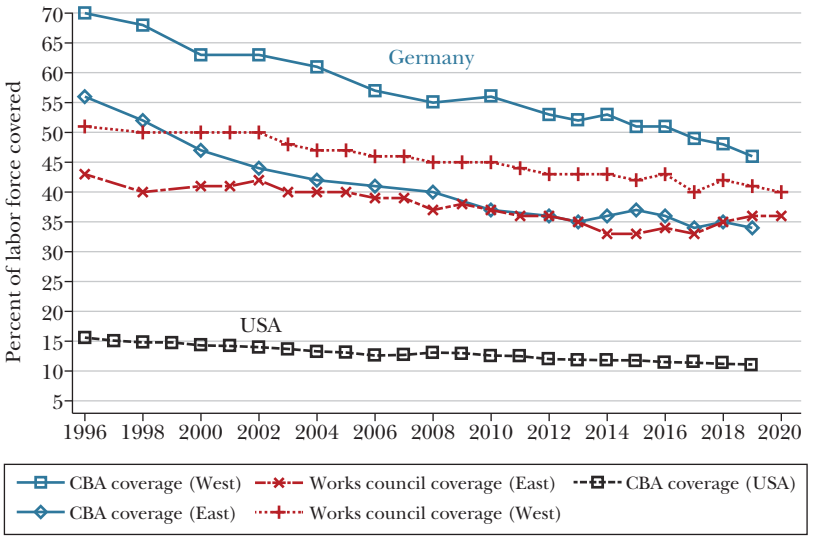
In this article, we present a primer on the "German model" of industrial relations. We organize our paper along its two key pillars. The first pillar is the sectoral bargaining system. In Germany, unions and employer associations engage in bargaining at the industry-region level, leading to broader coverage than in the United States. Meanwhile, partial decentralization of bargaining to the firm level—through flexibility provisions in sectoral agreements, or direct negotiations between individual firms and sectoral unions—gives firms some space to adapt to changing circumstances. However, this flexibility has also resulted in a gradual erosion of bargaining coverage. The second pillar of the German model is firm-level codetermination. Workers are integrated into corporate decision-making through membership on company boards and the formation of "works councils," leading to ongoing cooperative dialogue between shareholders, managers, and workers.

Overall, the German model combines centralized "social partnership" between unions and employer associations at the industry-region level with decentralized mechanisms for local wage-setting, dialogue, and customization of employment conditions.

The US industrial relations system is starkly different. American firms are run by managers on behalf of shareholders, within a legal structure that effectively bans cooperative forms of institutionalized worker voice akin to codetermination, in pursuit of "unencumbered" managerial decision-making (Harlin 1982).[2] US collective bargaining occurs exclusively at the bargaining unit or establishment level—rather than at the sectoral level—thereby giving individual employers strong incentives to resist unionization. Unionization elections are highly contentious and successful unionization is associated with lower profits and establishment closures (Lee and Mas 2012; Frandsen 2021; Wang and Young 2022). Over the past few

[2]These legal provisions were historically designed to ban employer-dominated unions (for example, see §8(a)(2) of the National Labor Relations Act), and subsequent judicial decisions have further narrowed the scope of unions' bargaining rights.

Figure 2
Institutional Coverage over Time in Germany and the United States



Source: Ellguth and Kohaut (2020) for the German numbers; Visser (2021) and OECD (2021) for the US numbers.
Note: This figure shows collective bargaining coverage in East and West Germany (blue lines) and the United States (black line), as well as works council coverage in East and West Germany (red lines) from 1996 to 2020. CBA stands for "collective bargaining agreement."

decades, private-sector collective bargaining coverage has been almost completely eroded (Farber et al. 2021).

A recurrent theme in our discussion of the German model will be a tension at the heart of the model: between firms' flexibility and workers' collective bargaining strength. Since the 1990s, the German model has become more decentralized and flexible. This evolution has arguably contributed to reductions in unemployment and increases in economic growth, but it has also entailed a substantial erosion of collective bargaining and works council coverage (as Figure 2 illustrates) and a weakening of bargaining agreements. This erosion may explain Germany's slowly increasing—and perhaps underappreciated—exposure to the afflictions suffered by other developed-world labor markets: rising wage inequality and the spread of low-wage, precarious jobs.

Sectoral Collective Bargaining

The German labor market is shaped by large-scale collective bargaining agreements containing schedules of minimum requirements for wages, hours, working conditions, entitlements, and promotion criteria for workers in different industries, regions, and occupations, and with different levels of skill and experience.

These agreements, typically negotiated at the industry-region level,[3] have broad coverage and create significant standardization in wages and working conditions—a sharp contrast to the patchwork system of employer-dominated wage-setting, individual bargaining, and (rare) establishment-level union bargaining prevalent in the United States. At the same time, the collective bargaining system in Germany allows for an unusual degree of decentralization and flexibility in wage-setting relative to the more rigid bargaining systems of many of its European neighbors—and even makes it relatively easy for employers to avoid coverage altogether.

The Bargaining Parties: Unions, Employer Associations, and Firms

Figure 3 visualizes the system of collective bargaining between sectoral trade unions and industry-region employer associations. The left of Figure 3 shows the worker side, the right shows the employer side, and the center illustrates the typical industry-region agreements. As we describe later, collective bargaining agreements can also sometimes be concluded between unions and individual firms.

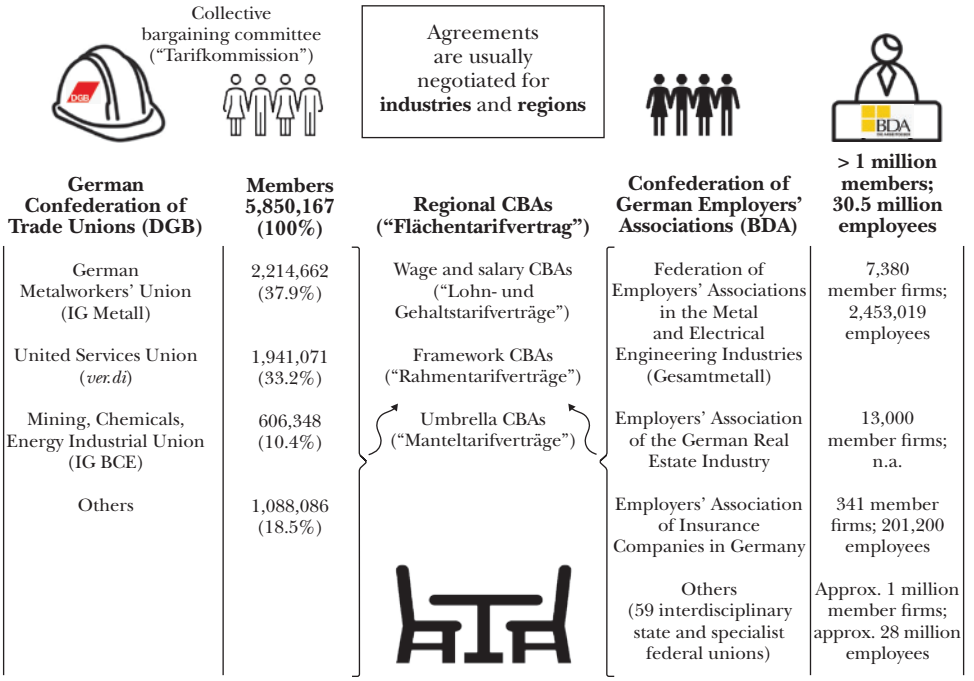
Unions. German unions are mostly organized at the sectoral level and belong to a small number of major trade union confederations. The most powerful confederation, the *Deutscher Gewerkschaftsbund* (DGB), oversees many of Germany's biggest unions, including *IG Metall* (manufacturing workers), *ver.di* (public-sector and services), *IG BCE* (mining and chemical industries), *GEW* (education and science), *IG BAU* (construction), and *NGG* (catering and restaurants) (DGB 2021). The DGB covers about 6 million workers; other union confederations include the *Deutscher Beamtenbund* (DBB), overseeing mainly civil service unions and covering about 1.3 million workers, and the Christian CGB, overseeing a variety of independent unions and covering about 300,000 workers (ETUI 2021). The union confederations compete for members and are differentiated by their political slant and attitudes toward collective bargaining. The DGB is mostly aligned with the center-left Social Democratic Party, though it maintains formal political neutrality and strives to always include a member of the center-right party (the Christian Democratic Union) on its governing staff. It remains strongly committed to broad sectoral bargaining. The DBB and CGB, by contrast, lean more toward the Christian Democrats and are less committed to industry-level bargaining—for example, the DBB contains several member unions organized at the level of granular occupations rather than industries.

German unions enjoy widespread public support and trust—about 73 percent of Germans agree that "workers need strong unions," compared to 49 percent of Americans.[4] The partisan gap in support for unions is also much smaller in Germany

[3] The industry-region bargaining system has its origins in the Stinnes-Legien Agreement and Collective Agreements Order of 1918, negotiated between moderate trade unions and major industry leaders against the backdrop of an unstable post–World War I provisional government and the threat of violent revolution from radical worker movements (Winkler 1998; Silvia 2013). The agreements institutionalized collective bargaining at the industry-region level, because this was the natural intersection between existing industry-region employer associations and industry-level trade unions. German employers had been organizing in local cartels throughout the nineteenth century, while trade unions formed at the industry level (Lepinski 1959; Silvia 2013).

[4] Authors' calculations using the International Social Survey Panel (ISSP Research Group 2017).

Figure 3
The German Bargaining Framework



Source: Deutscher Gewerkschaftsbund (2021) and BDA Die Arbeitgeber (2022).
Note: This figure illustrates the structure of collective bargaining between unions (left-hand side) and employer associations (right-hand side), with the center illustrating typical industry-region agreements. CBA stands for "collective bargaining agreement."

than in the United States.[5] The Christian Democrats (center-right party) are broadly supportive of collective bargaining: the party's election manifesto (Bundestagswahl 2021) asserted that the "social partnership" between unions and employer associations is at the core of Germany's economic success, called for more sectoral bargaining in the EU, and declared an intention to legislatively extend a greater number of bargaining agreements (see below for a discussion of these extensions).

German unions are prominent in public discourse, and often engage in policy lobbying. For example, they were instrumental in campaigning for the introduction of a federal minimum wage in 2015. They also fund research centers and think tanks, most notably the DGB-affiliated Hans Böckler Foundation. Trade union research institutes (and rival research institutes sponsored by employer associations) play a major role in economic policy discussions, frequently appearing in the media or writing widely covered reports.

[5] For example, 83 percent of 2013 Social Democratic Party (center-left) voters and 68 percent of Christian Democratic Union (center-right) voters agree with the statement in the text, compared with 63 percent of 2012 Obama voters and 26 percent of Romney voters.

Employer Associations. German employers organize in associations at the industry-region level, similarly to workers, with these associations in turn belonging to umbrella employer federations ultimately organized in the Federal Society of the German Employer Associations (*Bundesvereinigung der Deutschen Arbeitgeberverbände*, BDA). The BDA comprises 14 interdisciplinary regional associations and 48 federal umbrella associations representing more than 6,500 individual employer associations. Among the largest and most powerful umbrella associations are the metal and electrical industry association *Gesamtmetall*, the insurance companies association *Arbeitgeberverband der Versicherungsunternehmen in Deutschland*, and the chemical industry association *Bundesarbeitgeberverband Chemie*, as illustrated in Figure 3.

The primary function of employer associations is to engage in coordinated collective bargaining, but like unions, they also have several auxiliary functions. First, they engage in business lobbying focused on labor market policy, complementing the lobbying efforts of trade associations (*Wirtschaftsverbände*) and chambers of commerce and industry (*Industrie und Handelskammern*). For example, they have campaigned against wealth and inheritance taxes and lobbied for the abolition of a tax on high-earning individuals and firms (BDI 2021; BDA 2021). Second, employer associations often provide member companies with additional benefits such as legal advice and strike insurance. Third, they fund research institutes that play a major role in public discourse and also engage in lobbying and advocacy. Finally, employer associations are prominent networking hubs in the business world (Silvia and Schroeder 2007).

Relationships between Unions and Employer Associations. Interactions between major employer associations and major trade unions in Germany tend to be adversarial but respectful. The DGB and BDA are protective of their status as the economy's defining "social partners," and take pride in the industrial peace and low levels of strikes accompanying their partnership: for example, Germany lost only 5 days per 1,000 employees to strikes between 2001 and 2007, compared to 30 days in the United States (Lesch 2009). Both the DGB and BDA are wary of fragmentation of the industrial relations system; for example, they jointly lobbied for the introduction of a 2015 "unity law" declaring that only the largest bargaining agreement (in share of unionized workers) in an establishment could apply to that establishment. The law was intended to undercut a proliferation of occupation-specific unions representing highly skilled or hard-to-replace workers such as train conductors (*Gewerkschaft Deutscher Lokomotivführer*), who were demanding large wage increases and threatening to strike. Employer associations disliked the high wage demands and threat of strikes; the DGB argued that the demands of specific worker groups would exacerbate inequality and undermine the solidaristic principle of moderating wages at the top in order to boost wages at the bottom. The DGB was perhaps also motivated by an opportunity to consolidate its own power (Behrens 2016).[6]

International Perspective. Comparing Germany to the United States, it is tempting to draw analogies between, for example, the DGB in Germany and the AFL-CIO in

[6]Several large unions, such as *ver.di*, opposed and even unsuccessfully challenged the law before Germany's Federal Constitutional Court, asserting that it curtailed the freedom of association and individual unions' rights to strike (1 BvR 1571/15 -, Rn. 1-24, 1 BvR 571/16 -, Rn. 1-23).

the United States, or the BDA and the US Chamber of Commerce. However, there are several crucial differences.

First, membership is the lifeblood of American unions, whose influence is directly determined by the share of workers who have voted to collectivize their workplace and join a union. By contrast, membership matters only indirectly for German unions, because (as we describe below) bargaining happens at the sectoral level and agreements apply to both unionized and non-unionized workers in participating firms. To a German worker, joining a union is closer to becoming a fee-paying member of a political party. This has a few implications. It means German unions (mostly) do not have to engage in conflictual employer-specific unionization drives, which may help explain their enduring and bipartisan popularity. It also means that German unions have remained strong in the private sector even as union membership has diminished, whereas in the United States, diminishing membership has devastated private-sector union influence.

Second, American employer associations like the Chamber of Commerce traditionally adopt an actively hostile stance toward organized labor. By contrast, engaging in collective bargaining is the *raison d'être* of German employer associations, which derive their public legitimacy and membership appeal from their status as "social partners" with the unions. They are therefore highly tolerant of organized labor.

The German model of collective bargaining is also distinct from other European countries with strong unions. Countries such as Sweden, Norway, France, and Italy also feature two to four large, competing, widely legitimate union confederations paired with a large employer confederation in a stylized "social partnership" built around sectoral bargaining. However, this structural similarity obscures several axes of heterogeneity. First, unlike the Nordic countries and France—which have a "tripartite" industrial relations system where the government plays an active role—the German government is largely excluded from the industrial relations system.[7] Second, unlike in Denmark, Sweden, and Finland, where unions directly administer social insurance, in Germany social insurance is handled by the government, with no direct role for unions. Third, while most Western European countries maintain a notion of "social partnership" between unions and employer associations, the national ideology of cooperative partnership appears historically strongest in Germany and the Nordic countries.

[7]The principle of bargaining autonomy (*Tarifautonomie*) bans the government from intervening in collective negotiations; this rule dates back to the collective agreements of 1918, which were negotiated under a temporary provisional government while the future of the German state was highly uncertain. The federal government has occasionally experimented with soft-touch tripartism during crises, inviting employer associations and unions to roundtable discussions with legislators. This happened during the 1960s, when the government attempted to organize macroeconomic Keynesian coordination with the bargaining parties; in July 2022, the government convened talks under the same name (*Konzertierte Aktion*) in an attempt to tame rising inflation (*Deutsche Welle* 2022). This also happened in the 1990s (the "Joint Initiative for More Jobs in Eastern Germany"), when the government tried to tackle high unemployment and sluggish growth in East Germany by encouraging flexibility provisions and attentiveness to employment effects in bargaining agreements (Eurofound 1997).

The Contents of Collective Agreements

German unions and employer associations negotiate a range of industry-region-level collective agreements (*Flächentarifvertrag*), which are differentiated by the topics they cover, as shown in Figure 3. First, wage and salary agreements (*Lohn- und Gehaltstarifverträge*), usually renegotiated on an annual or biannual basis, specify wage and salary floors for workers in the industry-region, often by occupational, skill, and experience group. The "favorability principle" (*Günstigkeitsprinzip*) allows employers to offer higher salaries or better working conditions than those stipulated in wage and salary agreements. Second, longer-running framework agreements (*Rahmentarifverträge*) define criteria for assigning workers or positions to salary groups. Finally, umbrella agreements (*Manteltarifverträge*) regulate general working conditions, including termination rules, vacation duration, sick leave, and overtime, and are typically in place over longer periods. There are a huge number of active collective bargaining agreements at any given moment—82,000 in 2021 (Schulten et al. 2021).

As one example, a 2021 framework agreement between the metalworkers' union (*IG Metall*) and the corresponding regional employer association (*Südwestmetall*) regulates how workers in the metal and electronics industry in the German state of Baden-Württemberg are assigned to salary groups. It defines a points system, with points assigned for a worker's education and experience as well as the complexity and autonomy of the worker's job.[8] A separately negotiated collective bargaining agreement then stipulates wage floors for each points group.

Although collective bargaining agreements are typically negotiated at the industry-region level, there is substantial coordination in bargaining behavior across regions. Representatives of a national union confederation or umbrella employer organization are usually involved in guiding negotiations in a pilot region, and other regions then often imitate the agreement reached in the pilot region, deviating to match local conditions.

The collective bargaining structure allows for flexible firm- or establishment-level bargaining in a few circumstances. First, some (typically very large) individual employers negotiate separate firm level agreements with the relevant union (*Firmentarifverträge* and *Haustarifverträge/Werkstarifverträge*). For example, *RAFI GmbH & Co. KG*, an electronics manufacturer of human-machine interface technology, concluded a 2020 agreement with the relevant union (*IG Metall*) which binds *RAFI* to the conditions of the pertinent industry-region-level agreements, including the one described above, *and* specifies several additional provisions for *RAFI* employees, including bonus payments and sabbaticals (Wochenblatt-Online 2020). As this example illustrates, the main function of firm-level bargaining agreements is to bind large, productive firms to *even higher* standards than those stipulated in industry-level agreements. (Of course, firms can also voluntarily pay above the wage floors without such formal agreements, as discussed above.)

[8] See *Entgeltrahmen-Tarifvertrag für Beschäftigte in der Metall- und Elektroindustrie in Baden-Württemberg* (Salary Framework Collective Bargaining Agreement for Employees in the Metal and Electronics Industry in Baden-Württemberg) (IG Metall Bezirk Baden-Württemberg 2021).

Second, "hardship" and "opening" clauses, which are included in some collective bargaining agreements, allow firms to negotiate agreements (*Betriebsvereinbarung*) with their workforce that involve deviations *below* the wage, hour, or amenity requirements imposed by the industry-region agreements.

Hardship clauses apply only to firms or establishments in severe financial distress, and negotiations under these clauses involve, for example, workers agreeing to delay the implementation of collectively bargained wage increases until the financial situation of the establishment improves, or agreeing to temporary wage and hour cuts to prevent layoffs (Rehder 2003; Seifert and Massa-Wirth 2005). Dubbed "employment pacts," the latter kind of agreement likely played a role in preventing layoffs during the Great Recession of 2008–2009, although government-administered "short-time work" policies were the most important lever blunting the employment impacts of the crisis (Burda and Hunt 2011).

Meanwhile, *opening* clauses allow the negotiation of permanent employer-specific deviations from bargaining agreements. The criteria for using an opening clause vary; as one illustrative example, general opening clauses included in bargaining agreements in the metal industry since the mid-2000s allow companies to make deviations that "secure employment and create new jobs" or "[improve] competitiveness, innovative capability, and investment conditions" (Schulten and Bispinck 2018).

To use a hardship or opening clause, an employer typically negotiates an agreement with its works council (a shop-floor codetermination institution we cover in greater depth in the next section) and then submits the agreement to the sectoral union and employer association for approval.[9]

The relatively widespread use of hardship and opening clauses is unique to Germany and is one source of the unusual flexibility of the German system, on top of the state-level regional differentiation built into the bargaining system (compared to national sectoral bargaining as in, for example, Italy). In other countries, the scope of any such firm-level deviations from sectoral bargaining agreements is typically tightly circumscribed (as in France, for example), or they are simply less common (ETUI 2021).

Collective Bargaining Coverage

A bargaining agreement negotiated between a German union and an employer association covers all firms belonging to the signatory employer association. Covered firms typically extend coverage to all employees, regardless of union membership.[10]

[9] However, even employers who report not having a works council appear to use opening clauses with similar frequency to employers with a council (authors' calculations using the IAB establishment panel; Bellmann et al. 2021). We do not know of research reconciling this empirical pattern with the conventional wisdom we describe above.

[10] The law in principle allows for discrimination by union membership (for example, BAG 4 AZR 64/08). The law does not prohibit firms from granting coverage to all employees (BAG 4 AZR 366/09), a route firms typically take to reduce individual employees' incentives to unionize. An important exception is high-paid jobs, like managers or senior engineers with individually negotiated, above-collective-agreement salaries and working conditions (*Außertariflicher Arbeitsvertrag*); bargaining agreements often leave out these jobs.

Thus, although only 15 percent of German workers belong to a union (ETUI 2021), about 52 percent work in establishments covered by a collective bargaining agreement (Ellguth and Kohaut 2020). This stands in striking contrast to the United States, which had a private sector unionization rate of 6.1 percent in 2021 (Bureau of Labor Statistics 2022), and which has no capacity for bargaining coverage to substantially exceed unionization rates. This decoupling of coverage and union membership means German unions usually do not engage in employer-specific unionization drives as in the United States.

The German Labor Ministry can legislatively extend an agreement to cover *all* firms in the relevant industry-region (not just those belonging to the signatory employer association), if supported by a committee composed of union and employer representatives (*Tarifausschuss*). The threat of legislative extension was historically used to deter firms from exiting bargaining agreements en masse. These extensions were often supported by high-wage employers who wished to raise rivals' costs (Haucap, Pauly, and Wey 1999). However, extensions have become somewhat less common over time (Müller and Schulten 2019).

Why Do German Employers Opt into Coverage?

In contrast to the United States, where bargaining coverage is determined by whether *workers* choose to unionize, in Germany, individual *employers* opt in or out of coverage by industry-region collective bargaining agreements by joining or leaving the signatory employer association.[11] A growing number of employer associations even allow membership without participation in the relevant collective bargaining agreements (*OT-Mitgliedschaft*, see Behrens and Helfen 2016). This voluntary participation is a defining feature of the German model and the second contributor to its flexibility—and to the recent deterioration in bargaining coverage, as we discuss below. By comparison, in countries like France or Sweden, coverage is essentially mandatory and hence much higher, either due to frequent legislative extensions, as in France, or near-universal union membership and pressure to join agreements, as in Sweden (ETUI 2021).

Why do German employers ever join employer associations, thereby restricting their wage-setting discretion? First, membership in an association guarantees employers access to peaceful, coordinated, and widely legitimate mechanisms of dispute resolution through sectoral bargaining. In fact, active collective bargaining agreements preclude unions from strikes pertaining to any matters regulated in the pertinent collective bargaining agreement (*Friedenspflicht*).

Second, membership brings various side benefits, including access to strike insurance, legal advice, lobbying support, and professional networking.

Third, employers—especially large ones—may face pressure to join from workers and sectoral unions. Tesla's 2022 expansion into Germany provides an illustrative example.[12] During the first half of 2022, a new Tesla factory near Berlin has

[11] In the latter case, existing collective bargaining agreements remain active until expiry for incumbent workers (§3 (3) TVG).

[12] However, active exercises of union power like this are relatively uncommon, and Germany, unlike the United States, does not regularly see acrimonious conflicts over collective bargaining in major firms.

faced several complaints over its wage policies. In particular, wages are low relative to nearby manufacturing firms covered by sectoral agreements (Raymunt 2022), and Tesla has begun raising the wages offered to new hires in an effort to increase recruitment, which has introduced a wage gap between new recruits and identically qualified incumbents (*Der Spiegel* 2022). Discontented workers have appealed to the local *IG Metall* (manufacturing union) branch, which has begun publicly agitating for Tesla to enter collective negotiations. The union suspects Tesla may try to fend off the pressure by offering a local wage agreement to the plant's works council (which, unlike the sectoral union, cannot call a strike during wage negotiations). *IG Metall* has also rebuffed Tesla's instruction to all Tesla employees to return to work in-person, stating: "In Germany an employer cannot dictate the rules just as he likes . . . Whoever does not agree with such one-sided demands and wants to stand against them has the power of unions behind them" (as reported by Kay 2022).

Facts about German Bargaining Coverage

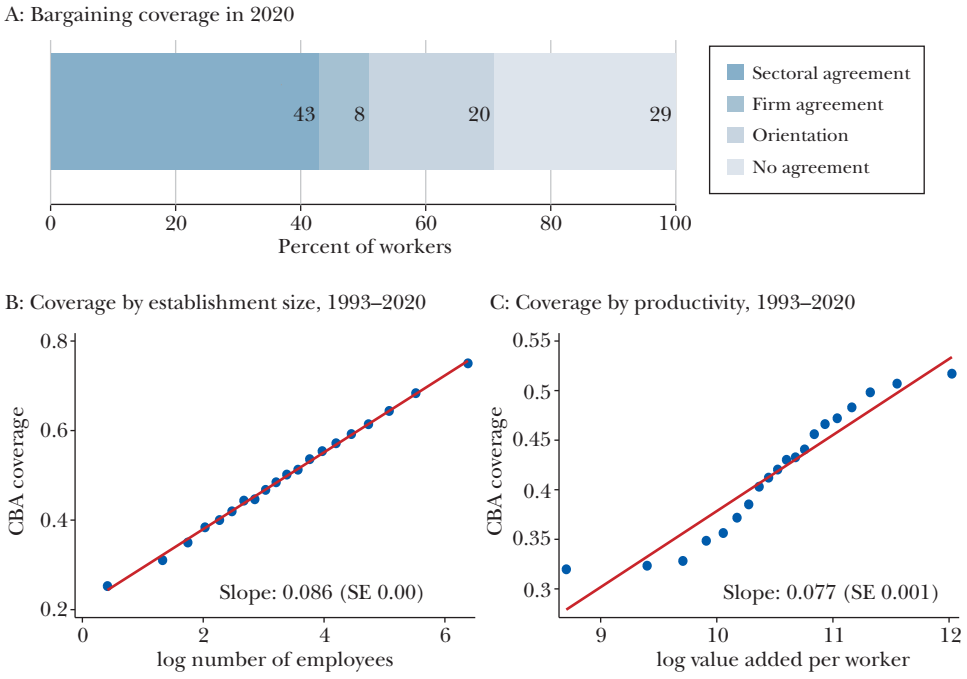
As of 2020, 27 percent of German establishments employing 52 percent of German workers are covered by a collective bargaining agreement, as shown in panel A of Figure 4 (Ellguth and Kohaut 2020). In particular, 43 percent of workers are covered by a sectoral agreement and 8 percent by firm-specific agreements. A further 20 percent of workers are employed by establishments reporting an "orientation" toward a bargaining agreement, meaning that they informally imitate the relevant agreement's prescribed wages and working conditions, but retain discretion to deviate from those prescriptions. This leaves 29 percent of German workers who are not covered, explicitly or by imitation, by a bargaining agreement.

Formal bargaining coverage in Germany is hence fairly high—substantially exceeding American union coverage even at the latter's mid-twentieth-century peak (Farber et al. 2021)—though significantly lower than coverage rates achieved through national bargaining or legislative extensions in countries like Sweden or France (as shown earlier in Figure 1).

Panel B of Figure 4 shows that coverage rates are strongly increasing in establishment size, reflecting the higher propensity of larger firms to join employer associations: only 10–20 percent of establishments with under 100 employees are covered by a collective bargaining agreement, compared to 50–60 percent of establishments with more than 500 employees. Larger firms are more likely to join employer associations for several reasons: they tend to be more productive and are hence more likely to pay high wages anyway, they may benefit more from the non-bargaining functions of employer associations (like lobbying), and unions tend to focus their pressure on large firms. Indeed, panel C of Figure 4 shows that coverage rates are also higher among more productive firms (by value added per worker), supporting the hypothesis that some firms join employer associations because they would have paid high wages anyway.

Erosion and Decentralization. The aforementioned statistics for 2020 reflect a steep drop in German bargaining coverage since the mid-1990s, when about 70 percent of German workers were covered, as shown in Figure 2. Employer association membership (and hence coverage by a collective bargaining agreement)

Figure 4
Collective Bargaining Coverage



Source: Panel A based on Ellguth and Kohaut (2020). Panels B and C based on IAB Establishment Panel (Umkehrer 2017; Bellmann et al. 2021); authors' own calculations.
Note: Panel A illustrates the share of German workers in 2020 covered by a sectoral bargaining agreement, a company-level bargaining agreement, or an informal orientation toward a collective bargaining agreement. Panels B and C plot establishment-level regressions of a dummy for being covered by a bargaining agreement on log number of employees (panel B) or log value added per worker (panel C), controlling for year dummies interacted with an East/West Germany dummy, and three-digit industry dummies.

has declined substantially, especially among small- and medium-sized firms (Hassel 1999; Bispinck and Schulten 2010; Kügler, Schönberg, and Schreiner 2018). Informal "orientation" toward collective bargaining agreements has grown over the same period (Oberfichtner and Schnabel 2018). Among covered firms, the proliferation of general "opening clauses" since the mid-2000s have allowed firms to negotiate deviations below the floors set by a collective bargaining agreement. Representative data on opening clauses is scant and at times conflicting. Based on a 2015 survey of works council members, 21 percent of establishments with at least 20 employees (and a works council) made use of opening clauses—for example, to pay wages below the level set by the collective bargaining agreement (Amlinger and Bispinck 2016)—and data from the IAB Establishment Panel show a substantially higher prevalence of opening clauses (Boeri et al. 2021). Finally, large and high-paying firms have increasingly evaded collective bargaining agreements for their lowest-paid workers by outsourcing jobs to uncovered supplier firms. For example, the proportion of retail establishments with a cleaning worker on their

own staff declined from 82 percent in 1975 to 20 percent in 2009 (Goldschmidt and Schmieder 2017), reflecting a surge in outsourcing of food, cleaning, security, and logistics jobs in the economy.

The sources of the erosion and decentralization of German collective bargaining since the 1990s remain an active area of debate; here, we name the main candidates.

First, increasing exposure to foreign competition and a prolonged recession in the 1990s drove many German firms—especially small ones—into financial distress and provoked a flight from employer associations to avoid wage floors of collective bargaining agreements (Silvia and Schroeder 2007; Dustmann et al. 2014; Raess 2014).

Second, the dissolution of the Soviet Union allowed employers to more credibly threaten outsourcing production to low-wage Eastern European neighbors (Dauth, Findeisen, and Suedekum 2014), giving employer associations greater bargaining power and allowing them to lobby for opening clauses and other flexibility provisions in collective bargaining agreements. Unions also began to embrace opening clauses and to negotiate firm-level "employment pacts" to protect against the growing threat of layoffs (Schulten and Bispinck 2018).

Third, beginning in the 1980s, small, unproductive employers could not keep up with the wage floors negotiated by employer associations dominated by large, highly productive firms, and hence exited the associations (Silvia 1997; Dustmann, Ludsteck, and Schönberg 2009).

While similar factors have been linked to the decline of collective bargaining coverage in the United States (Acemoglu, Aghion, and Violante 2001; Farber and Western 2001), the collapse of US unions was also partially driven by political and legislative changes (including growing employer hostility to unions, the rise of the shareholder value paradigm, and the spread of right-to-work laws). By contrast, in Germany, the basic consensus in favor of the industrial relations regime has remained solid since the 1950s. Moreover, although some of the changes since the 1990s were informally encouraged by the government (for example, through the Joint Initiative for More Jobs in Eastern Germany mentioned in footnote 7), they were not implemented through legislative reforms.

The decline in German collective bargaining coverage shows no sign of abating, as coverage has kept dropping in each new cohort of firms (Card, Heining, and Kline 2013). However, growing inequality and the expansion of a nascent low-wage sector unconstrained by collective bargaining agreements has also motivated pushback against the erosion and decentralization of collective wage-setting. First, in 2015, following a successful union campaign, the government introduced Germany's first federal minimum wage (Dustmann et al. 2022). Second, as mentioned above, unions and employer associations lobbied successfully for a collective bargaining agreement "unity law" in 2015, in an attempt to arrest the gradual fragmentation of bargaining in specific sectors. Third, political parties have declared intentions to mandate broader coverage. For instance, legislative extensions of bargaining agreements are on the table (Soziale Politik für Dich 2017, 2021; Bundestagswahl 2021), and the 2021 *Gesundheitsversorgungsweiterentwicklungsgesetz* (Health Care Advancement Act) will restrict public payments to only those long-term care providers that pay wages compliant with collective agreements.

Evidence on the Effects of the German Sectoral Bargaining System

What are the effects of German wage-setting institutions on employment levels and the wage structure? While causal estimates of these effects are scarce, we review the existing evidence that speaks to this question.

Zooming out to the aggregate time series suggests that the erosion and decentralization of collective bargaining since the 1990s weakened an institution that had previously held up wages at the bottom, constrained wage inequality, and increased unemployment by restricting firm-level wage setting. Several patterns support this account. First, beginning in the 1990s, real wages have declined in the lower deciles of the German wage distribution (Dustmann, Ludsteck, and Schönberg 2009). Second, earnings inequality has risen dramatically, with about 25 percent of the increase driven by growing heterogeneity in pay across firms (Card, Heining, and Kline 2013).[13] Third, the German economy experienced a remarkable resurgence beginning in the mid-2000s, with the unemployment rate dropping from about 10 percent to below 5 percent, potentially due to increased competitiveness of manufacturing exporters thanks to lower real wages at the bottom of the distribution (Dustmann et al. 2014).[14]

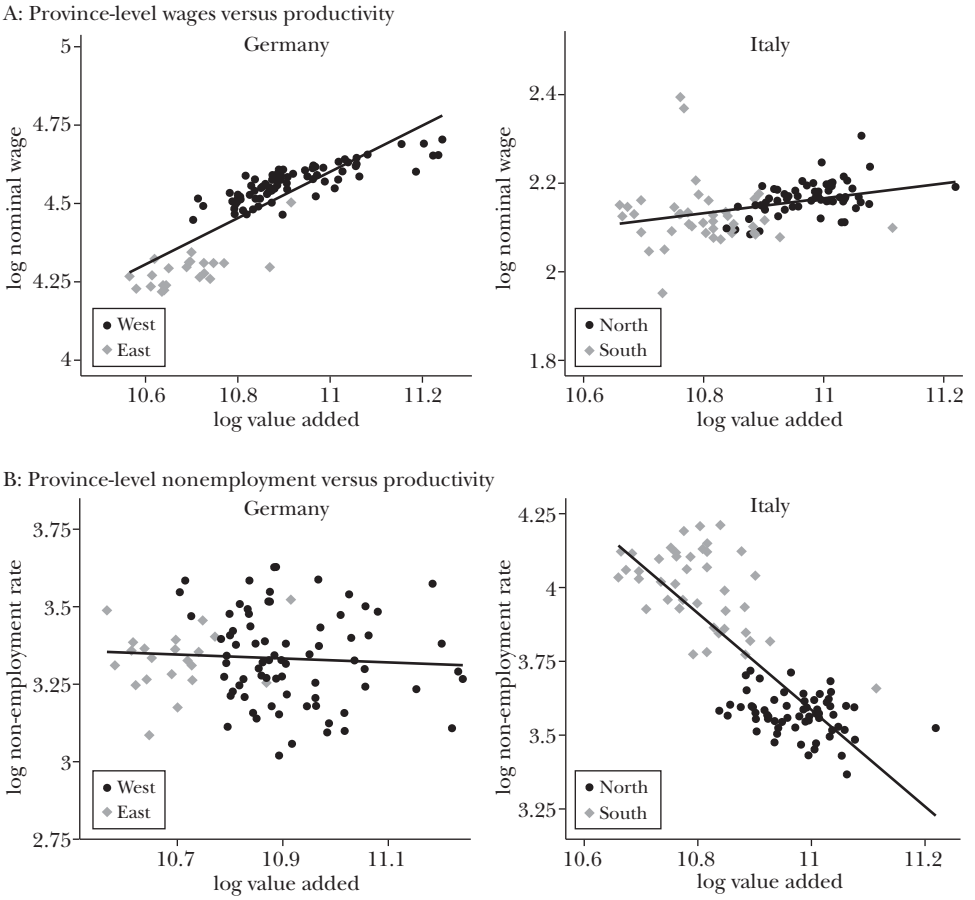
One inherent limitation of this time series perspective is the presence of other contemporaneous trends, like globalization and skill-biased technological change. (In fact, the erosion of collective bargaining may itself have been an outcome of these forces, as discussed in Acemoglu, Aghion, and Violante 2001.) Such time series narratives also tend to be quite flexible: in the 1950s–1980s, Germany's strong performance was attributed to the bright side of sectoral bargaining (Silvia 1997), a narrative that flipped in the 1990s (Ochel 2005; Schulten and Bispinck 2018) and has been changing again following Germany's success since the late 2000s.

Cross-sectional international comparisons paint a picture similar to the time-series narrative. As reproduced in Figure 5, a striking figure by Boeri et al. (2021) shows that the German system—thanks to non-mandatory employer participation, the regionalization of sectoral bargaining, and the spread of opening clauses—allows wages to vary according to regional productivity and hence maintains high employment rates everywhere, even in lower-productivity areas, particularly East Germany. By contrast, the Italian system—which imposes uniform wage floors across all regions with limited local wage adjustments—largely delinks wages from regional productivity and hence depresses employment in low-productivity regions, such as Southern Italy. Again, these results are consistent with claims that the more rigid twentieth-century German bargaining system compressed wages at the expense of elevated unemployment, and reforms to the bargaining system since the 1990s have resulted in greater wage dispersion but increased employment.

[13] Hirsch and Mueller (2020) provide evidence that this increase in dispersion of firm pay premia is partially explained by declining bargaining coverage.

[14] An important alternative hypothesis for this resurgence points to the Hartz reforms of the early 2000s. These reforms were the closest German analogue to the Reagan/Thatcher reforms of the 1980s. They cut the generosity of unemployment benefits and reformed active labor market policies (for discussion, see Krebs and Scheffel 2013; Price 2018; Hartung, Jung, and Kuhn 2018).

Figure 5
Collective Bargaining Flexibility in Germany and Italy



Source: This figure reproduces Figures 4 and 6 of Boeri et al. (2021).
Note: Panels A and B show scatterplots (each dot representing a province) of mean log wages against mean log value added, separately for Germany and Italy. Panels C and D show province-level scatterplots of log nonemployment against mean log value added, again separately for Germany and Italy. The distinction between West/East Germany is analogous to the distinction between North/South Italy, in that the former region tends to be wealthier and more productive in each case. Data are from 2010.

Firm-level evidence also suggests that the contemporary collective bargaining system slightly raises mean wages in covered firms, compresses within-firm wage distributions, and raises the average proportion of rents shared with workers (while reducing firm-level wage-setting discretion at the margin). More specifically, *uncontrolled* cross-sectional comparisons of firms covered and uncovered by sectoral bargaining indicate 10–30 percent higher average wages in covered firms (Dustmann and Schönberg 2009; Addison et al. 2016). However, controlling for worker and firm characteristics reduces this premium to about 2 percent (Hirsch and Mueller 2020), with event studies of firms exiting and entering agreements

suggesting a 3–4 percent premium (Addison, Teixeira, Evers, and Bellmann 2014, 2016). Increases in profits or productivity are passed on to workers *less* so in covered firms (Gürtzgen 2009), but wages conditional on (static) rents are higher (Hirsch and Mueller 2020). Meanwhile, within-firm wage dispersion tends to be lower in covered firms (Dustmann and Schönberg 2009), which also invest more in apprenticeship training in line with theories of wage compression and training provision (as in Acemoglu and Pischke 1999).

As coverage is largely voluntary for firms in Germany, these firm-level comparisons need not only reflect the causal effect of coverage. Firms opting into collective agreements might pay high and compressed wages anyway. There is no existing source of identification for calculating the union wage premium in Germany mirroring the close union election regression discontinuity design in the United States (DiNardo and Lee 2004) or sharp policy variation as in Portugal (Hijzen and Martins 2020). More broadly, even an ideal firm-level experiment would leave open the question of equilibrium effects of sectoral bargaining through market spillovers or norms and expectations about pay, as suggested by the phenomenon of "orientation" to collective bargaining agreement wages by uncovered firms (see also Falk, Fehr, and Zehnder 2006; Western and Rosenfeld 2011).

Sectoral Bargaining and the Quality of Industrial Relations

The collective bargaining system also plausibly contributes to Germany's remarkably harmonious industrial relations, which are built around the "social partnership" between union confederations and employer associations. It does so through two mechanisms.

First, Germany's system of sectoral bargaining elevates zero-sum bargaining over the division of rents to the higher level of industry-region negotiations, in contrast to the adversarial firm-level bargaining system in the United States (Moene, Wallerstein, and Hoel 1992). When negotiations do take place at the firm level in Germany (as under opening clauses), these negotiations still occur in the shadow of the industry-region agreements, as evidenced by the frequent requirement to submit deviations to the sectoral bargaining parties for approval, and by the increasing number of firms informally "orienting" their pay policies to collective bargaining agreements.

Second, employers' ability to opt in or out of collective bargaining coverage, and the decoupling of bargaining coverage from firm-level union membership, eliminates individual employers' incentives to crack down on union activity in the firm.

Codetermination

The second pillar of German industrial relations is codetermination (*Mitbestimmung*), which refers to the legally mandated integration of workers into corporate governance and decision-making. German codetermination comes in two forms: representation on corporate boards, and works councils. In the first form, workers

elect representatives to company boards, thereby gaining a vote in major decisions and the appointment, supervision, and dismissal of top corporate management. Board representation is restricted to relatively large firms, and is mandatory in those firms. In the second form, workers elect establishment- and firm-level works councils tasked with participating in day-to-day managerial decision-making. Workers have a right to form works councils in all firms except the very smallest ones, so that this second form of codetermination is more widespread. Corporate governance under codetermination contrasts with the American system of corporate governance, where boards are composed exclusively of shareholder representatives and executives, and day-to-day decision-making is purely in the hands of managers.

In this section, we describe how codetermination operates, illustrate its interaction with industry-level bargaining, review evidence on its effects, and connect it to the overall trends of erosion and decentralization in the German model.

Board Representation

Germany was the first country in the world to implement wide-scale board-level codetermination, with legal provisions for board-level representation going back to the Weimar Republic (Winkler 1998). Following World War II, the institution in its modern form was introduced by the British occupiers, who imposed "parity" codetermination requirements (50–50 shareholder-worker board representation) on firms in the iron, coal, and steel sectors, with the goal of breaking up the power of industry leaders who had helped drive both World Wars. Lobbying campaigns by German unions later led to the extension of the institution (in a substantially weaker form) to all sectors by legislation passed in 1952 and 1976 (for more historical background, see Jäger, Noy, and Schoefer 2022a). Other European countries later followed suit; however, board-level codetermination remains uncommon internationally, with fewer than 20 countries featuring the institution today (Jäger, Noy, and Schoefer 2022b).

In general, German firms with more than 500 employees must have worker representatives on their supervisory boards, alongside the regular shareholder representatives. A firm's *supervisory board* selects and oversees the firm's *executive board*, which is composed of senior executives and is tasked with day-to-day management. The supervisory board also participates in major decisions, such as decisions about large investments or significant changes to company operations.

Minority, Quasi-Parity, and Parity Representation. There are three tiers of board-level codetermination requirements, applying to different groups of German firms (Jäger, Schoefer, and Heining 2021; ETUI 2021). First, under *minority representation*, firms with between 500 and 2,000 employees (and stock corporations incorporated before August 1994, regardless of size) must appoint worker representatives to 33 percent of the seats on their supervisory board. In these firms, the worker representatives are company employees directly elected by workers. Second, under *quasi-parity representation*, firms with more than 2,000 employees must appoint worker representatives to 50 percent of the seats on their supervisory board, though shareholder representatives receive the tie-breaking vote. In these firms, some worker representatives are elected company employees, others are external representatives

of the union covering the company's workforce, and at least one is chosen by senior managers as a representative of their interests as employees. Finally, there exists true *parity representation*, where 50 percent of the seats go to workers, and shareholder representatives do *not* receive a tie-breaking vote (instead, a neutral chair, appointed by majorities of both the shareholder and worker representatives, holds the deciding vote). However, parity representation is limited to firms with more than 1,000 employees in the iron, coal, and steel sectors, as a political relic of the post–World War II arrangements.

Quasi-parity and parity representation are unique to Germany and are the strongest forms of board-level codetermination in the world; all other countries with board-level codetermination laws have implemented minority representation. Apart from this, the German board-level codetermination system is virtually identical to the systems present in many other European countries (Jäger, Noy, and Schoefer 2022b).

Labor in the Boardroom. Worker representatives have the same rights and obligations as shareholder representatives, and they can discuss and vote on any matter that comes before the supervisory board. In this way, workers have a direct voice in major strategic decisions. For example, in interviews, worker representatives describe lobbying for more generous pension plans, alerting shareholder representatives to job security and task duplication issues following a merger or acquisition, providing input on the construction of new company buildings, and pushing back against a focus on maximizing short-run returns (Gold, Kluge, and Conchon 2010, pp. 74, 84, 85, 94). They also describe collaborating with works councils or union representatives, to coordinate messaging or to lobby for legislative changes (Gold, Kluge, and Conchon 2010, pp. 76, 96, 97). Like shareholder representatives, worker representatives have a fiduciary duty to the company (rather than to workers), which leads to occasional tensions (for example, Gold, Kluge, and Conchon 2010, pp. 76, 77, 84).

Anecdotally, the relationships between worker and shareholder representatives on supervisory boards are friendly and collaborative. Most German executives are broadly supportive of board-level codetermination laws (Paster 2012), with some evidence that minority codetermination is viewed more favorably and quasi-parity codetermination is more likely to be opposed (Stettes 2007). Shareholder representatives appreciate the insights into workers' preferences and company operations provided by worker representatives (Gold, Kluge, and Conchon 2010, p. 93). Votes on supervisory boards are usually unanimous (Gold, Kluge, and Conchon 2010). Since worker representatives on their own cannot outvote shareholder majorities and recognize the importance of maintaining friendly and cooperative relationships, they are usually acquiescent and recognize the limits of their influence (see, for instance, Gold, Kluge, and Conchon 2010, pp. 74, 82).

The Effects of Board-Level Codetermination. The available quasi-experimental evidence suggests limited causal effects of board-level worker representation on, for example, wage-setting or investment, perhaps consistent with the limited power held by worker representatives owing to their minority vote share (Jäger,

Schoefer, and Heining 2021). A large literature using cross-sectional comparisons or simple regression-discontinuity designs similarly finds mixed effects (Conchon 2011). Firms also do not appear to bunch below the relevant size thresholds to avoid codetermination requirements (Jäger, Noy, and Schoefer 2022b), providing revealed-preference evidence that the institution does not harm firm performance enough to lead firms to distort their size to evade it.

According to anecdotal reports, board-level representatives may complement unions' and works councils' activities—for example, by sharing information gained from board meetings. Board-level codetermination has also been hypothesized to contribute to the environment of cooperation and social partnership that characterizes German industrial relations (Thelen 2014)—more on this below.

Board-Level Codetermination and the Erosion of the German Model. Setting aside a reform in 1994 (Jäger, Schoefer, and Heining 2021), board-level codetermination laws have remained largely untouched since the last major reform in 1976. There are no reliable statistics on the coverage of the institution, so it is difficult to tell whether it has experienced a decline in coverage over the past 30 years analogous to the decline experienced by collective bargaining agreements and works councils. There is anecdotal evidence of increasing attempts by large firms to evade board-level codetermination requirements through legal restructuring or by simply ignoring the mandates, suggesting that the gradual erosion of the German model may be affecting board-level representation as well (Sick 2020).

Works Councils

Works councils—the second facet of German codetermination—are committees of representatives elected by workers who have rights to participate in a variety of managerial decisions. They typically are a form of lower-level, "shop-floor" codetermination that complements board-level codetermination, although firms with multiple works councils across establishments also have a firm-level works council (*Gesamtbetriebsrat*). German works councils possess broader and stronger co-decision-making rights compared to board-level representatives—who, as we have noted, lack formal power due to their minority share on boards—and compared to other European countries' often anemic shop-floor codetermination institutions (Jäger, Noy, and Schoefer 2022b). The German works council system dates back more than a century, to the Stinnes-Legien Agreement of 1918 and the Works Council Act of 1920.

German law gives workers in any establishment with at least five employees the *right* (but not a requirement) to form a works council. If a works council is set up, the number of representatives on the council scales with the establishment's size, ranging from one in establishments with 5–20 workers to 15 in establishments with 1,001–1,500 workers (ETUI 2021). There are quotas for gender representation on councils. Responsibilities also scale with establishment size. In larger establishments, the works council sets up various subsidiary committees: a health and safety committee (in establishments with more than 50 workers), an economic committee (more than 100 workers) that scrutinizes company financials and is consulted on related matters, and a works committee (more than 200 workers) that deals with

day-to-day managerial issues. Additionally, in these larger establishments, some works council members are allowed to perform their duties full-time.

Almost 40 percent of German workers are covered by a works council (Ellguth and Kohaut 2020). These tend to be at larger establishments. As Figure 2 shows, these coverage numbers represent a moderate decline since the early 1990s, when about 50 percent of workers were covered by a works council (more on this below).

Works Council Powers. Works councils have a spectrum of powers in various areas (ETUI 2021). At the weaker end, they have various information and consultation rights: a right to be kept informed about the company's financial situation and a right to be consulted about planned changes that might affect workers, including changes to work methods, training, and health and safety procedures—though the employer usually has no obligation to follow their advice. At the stronger end, works councils can veto transfers, dismissals, or appointments of employees if they can show that the employer has acted unfairly or violated an existing agreement. Employers can appeal to a labor court to override the veto. At the very strongest end, works councils have full co-decision-making rights regarding working hours, vacations, workplace monitoring, bonuses and payment schemes, redundancy payments, and workplace amenities. In these areas, decisions must be jointly reached and approved by the employer and the works council, with both sides having the power to initiate proposals. However, works councils cannot initiate strikes. Disagreements are adjudicated by a conciliation committee (consisting of worker and employer representatives and chaired by a neutral arbiter who holds the deciding vote). Works councils are also responsible for the increasingly important job of negotiating over firm-level deviations from the requirements of the collective bargaining agreements.

These powers make German works councils among the strongest shop-floor codetermination institutions in the world, along with Swedish and Norwegian firm-level union representatives. While many other countries have shop-floor codetermination institutions, these institutions tend to grant workers narrow information, consultation, and arbitration rights, in contrast to the sweeping co-decision-making powers held by German works councils (Jäger, Noy, and Schoefer 2022b).

Works Councils and Unions. Although works councils were originally conceived as local representatives of industry-level trade unions, German law now maintains a clear legal separation between the two institutions (dating back to reforms in the 1950s aimed at weakening unions, as discussed in Jäger, Noy, and Schoefer 2022a). However, in practice, works council members frequently occupy leadership positions in unions, and unions are closely involved in the procedures to set up works councils. Works council elections even frequently feature political-party-style union lists, and councils engage in membership drives for unions (Behrens 2009). Councils are additionally formally tasked with monitoring compliance with collective bargaining agreements and employment regulations (§80 Works Council Act) and engage in negotiations under opening clauses.

The Effects of Works Councils. A long empirical literature compares firms and establishments with and without works councils, with the common finding that works councils are associated with slightly higher wages and productivity and more

compressed wage distributions (Addison 2009; Jirjahn and Smith 2018; Adam 2019; Hirsch and Mueller 2020; Schnabel 2020). However, the voluntary nature of works council coverage, the low employment threshold for workers' right to demand one, and an absence of compelling natural experiments make causal inference difficult. In general, it is plausible that works councils are more directly impactful than board-level codetermination, given that councillors are allocated a variety of direct decision-making powers that board-level representatives lack and interact more often with workers (at the shop floor). But due to a lack of sharp and exogenous variation, the effects of works councils on worker and firm outcomes remain an open research question.

Works Councils and the Erosion of the German Model. Works councils have played a dual role in the changes to German industrial relations since the 1990s. On the one hand, works councils have facilitated the partial decentralization of collective bargaining to the firm level, specifically the utilization of opening clauses. The associated negotiations under opening clauses have blurred the boundaries between cooperative codetermination and adversarial bargaining, and shifted Germany somewhat closer to the Nordic model, where establishment-level union representatives hold both codetermination and bargaining rights.

On the other hand, works councils have themselves been victims of the decline in collective institutions over the past three decades, as Figure 2 shows. The decline in coverage has been concentrated in medium-sized firms and, perhaps surprisingly, appears evident in all sectors (Addison et al. 2017). The causes of the decline are not well understood; one hypothesis is that it is part of a generalized decline in worker mobilization and the power of unions, and an increased willingness by managers and employers to avoid collective worker institutions.

Codetermination and the Quality of Industrial Relations

A longstanding hypothesis holds that Germany's codetermination institutions are partially responsible for its unusually harmonious industrial relations and culture of "social partnership" (for discussion, see Thelen 1991). By providing systematic opportunities for cooperation and conversation between employers and workers at the firm level while adversarial bargaining is outsourced to the sectoral level, codetermination might provide the foundations for friendlier partnership between firms and workers (Freeman and Lazear 1995). We do not know of compelling tests of this hypothesis. In cross-country event studies, Jäger, Noy, and Schoefer (2022b) find no evidence that codetermination reforms in European countries are associated with subsequent improvements in industrial relations, but their results have wide confidence intervals and only study incremental shifts in this single institution.

Conclusion

Overall, the contemporary German model shows that powerful unions, a relatively robust collective bargaining system, and involvement of workers in corporate decision-making are compatible with friendly and peaceful industrial relations and

with the avoidance of distortionary pitfalls traditionally thought to be associated with strong labor power. Several features of the model may underlie these outcomes: i) the outsourcing of most distributional conflicts to the industry-region (rather than firm) level; ii) the decoupling of bargaining coverage from workers' unionization status, which reduces employers' incentives to oppose unionization; and iii) the institutionalization of worker-management cooperation through codetermination. The result has been a long history of unusually harmonious industrial relations stretching back to the 1950s. Meanwhile, the (increasing) ease of nonparticipation in collective bargaining, the proliferation of opening clauses and other flexibility provisions, and the regionalization of bargaining mean that the contemporary German system seems much less likely to reduce employment, exclude potential labor market entrants, or slow down growth than sectoral bargaining systems in peer countries with more comprehensive and stricter coverage rules.

At the same time, the increasing flexibility of the German system means that Germany is no longer a poster child for strong sectoral bargaining. Bargaining coverage in Germany is middling, and decreasing. The flexibility to which Germany's strong macroeconomic performance is often attributed involves the omission of large segments of the labor market from bargaining coverage. Germany is now starting to face many of the challenges that its historically more rigid industrial relations system used to suppress: significant increases in earnings inequality, the spread of precarious work, and the gradual expansion of a low-wage sector that is now larger than the OECD average (though still 25 percent smaller than in the United States).

Frustration with these developments in Germany has led to the introduction of a more rigid national minimum wage and louder calls to strengthen both pillars of worker representation. The new, center-left government has proposed to extend collective bargaining coverage to more employers: for example, by formally extending more collective bargaining agreements and by making public procurement contingent on compliance with the relevant collective bargaining agreement (SPD, Gruene, and FDP 2021). Members of the governing coalition also plan to make it easier to prosecute employers who (illegally) oppose works council elections, to facilitate works councils for gig and platform workers, and to close loopholes that allow evasion of board-level codetermination (SPD, Gruene, and FDP 2021; *Handelsblatt* 2022).

The German model of industrial relations will continue to evolve as fault-lines that opened up in the 2000s continue to widen. The model will also shape and be shaped by new challenges. For instance, in response to the pandemic, collective bargaining agreements have started to include remote work provisions. "Crisis summits" between the bargaining partners and the government have discussed responses to issues like high inflation, an energy crisis precipitated by the Russian invasion of Ukraine, and gradual decarbonization of the economy (Kell 2022; *Deutsche Welle* 2022).

■ *We thank Alexander Busch, Jannis Hamida, Kilian Weil, and Pascal Zamorski for excellent research assistance. We thank Reinhard Bispinck and Steffen Müller for helpful comments. We thank Jörg Heining for help with the IAB Establishment Panel. Schoefer thanks the UC Berkeley Institute for Research on Labor and Employment for financial support; Jäger and Schoefer thank the National Science Foundation for support.*

References

Acemoglu, Daron, Philippe Aghion, and Giovanni Violante. 2001. "Deunionization, Technical Change and Inequality." *Carnegie-Rochester Conference Series on Public Policy* 55 (1): 229–64.

Acemoglu, Daron, and Jörn-Steffen Pischke. 1999. "The Structure of Wages and Investment in General Training." *Journal of Political Economy* 107 (3): 539–72.

Acemoglu, Daron, and Pascual Restrepo. 2020. "Robots and Jobs: Evidence from US Labor Markets." *Journal of Political Economy* 128 (6): 2188–244.

Adam, Julian. 2019. "Voluntary Quits: Do Works Councils Matter? An Analysis of the Reform of the German Works Constitution Act 2001." *Jahrbücher* für *Nationalökonomie und Statistik* 239 (1): 67–109.

Addison, John. 2009. *The Economics of Codetermination: Lessons from the German Experience*. New York: Palgrave Macmillan.

Addison, John, Paulino Teixeira, Katalin Evers, and Lutz Bellmann. 2014. "Indicative and Updated Estimates of the Collective Bargaining Premium in Germany." *Industrial Relations: A Journal of Economy and Society* 53 (1): 125–56.

Addison, John T., Paulino Teixeira, Katalin Evers, and Lutz Bellmann. 2016. "Is the Erosion Thesis Overblown? Alignment from Without in Germany." *Industrial Relations: A Journal of Economy and Society* 55 (3): 415–443.

Addison, John, Paulino Teixeira, André Pahnke, and Lutz Bellmann. 2017. "The Demise of a Model? The State of Collective Bargaining and Worker Representation in Germany." *Economic and Industrial Democracy* 38 (2): 193–234.

Amlinger, Marc, and Reinhard Bispinck. 2016. *Dezentralisierung der Tarifpolitik—Ergebnisse der WSI-Betriebsräte-befragung 2015*. Düsseldorf: WSI-Mitteilungen.

Anderson, Richard. 2012. "German Economic Strength: The Secrets of Success." *BBC*, August 16. https://www.bbc.com/ news/business-18868704.

Autor, David H. 2014. "Skills, Education, and the Rise of Earnings Inequality Among the 'Other 99 Percent.'" *Science* 344 (6186): 843–51.

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

BDA. 2021. *Die Arbeitgeber zur 20. Legislaturperiode*. Berlin: Bundesvereinigung der Deutschen Arbeitgeberverbände.

BDI. 2021. *Wahlprogramm-Check zur Bundestagswahl 2021*. Berlin: Federation of German Industries.

BDA Die Arbeitgeber. 2022. "Mitglieder." Bundesvereinigung der Deutschen Arbeitgeberverbände. https://arbeitgeber.de/en/mitglieder/ (accessed March 2, 2022).

Behrens, Martin. 2009. "Still Married after All These Years? Union Organizing and the Role of Works Councils in German Industrial Relations." *Industrial and Labor Relations Review* 62 (3): 275–93.

Behrens, Martin. 2016. "Employment Relations in Germany." In *Developments in German Industrial Relations*, edited by Ingrid Artus, Martin Behrens, Berndt Keller, Wenzel Matiaske, Werner Nienhüser, Britta Rehder, and Carsten Wirth, 1–30. Newcastle: Cambridge Scholars Publishing.

Behrens, Martin, and Markus Helfen. 2016. "Sachzwang oder Programm? Tarifpolitische Orientierungen und OT-Mitgliedschaft bei deutschen Arbeitgeberverbänden." *WSI-Mitteilungen* 69 (6): 452–59.

Bellmann, Lutz, Laura Brunner, Peter Ellguth, Philipp Grunau, Christian Hohendanner, Susanne Kohaut, Ute Leber, and et al. 2021. "IAB-Betriebspanel (IAB BP) – Version 9320 v1." doi:10.5164/IAB.IABBP9320.DE.EN.V1 (accessed April 4, 2022).

Bispinck, Reinhard, and Thorsten Schulten. 2010. *Sector-Level Bargaining and Possibilities for Deviations at Company-Level: The Case of Germany*. Dublin: Eurofound.

Boeri, Tito, Andrea Ichino, Enrico Moretti, and Johanna Posch. 2021. "Wage Equalization and Regional Misallocation: Evidence from Italian and German Provinces." *Journal of the European Economic Association* 19 (6): 3249–92.

Bundestagswahl. 2021. *Alle Wahlprogramme für die Bundestagswahl 2021*. https://www.bundestagswahl-2021.de/wahlprogramme/#cdu (accessed March 2, 2022).

Burda, Michael C., and Jennifer Hunt. 2011. "What Explains the German Labor Market Miracle in the Great Recession?" NBER Working Paper 17187.

Bureau of Labor Statistics (BLS). 2022. "Union Members—2021." News release no. USDL-22-0079, January 20, 2022, https://www.bls.gov/news.release/pdf/union2.pdf.

Campbell, Alexia Fernández. 2019. "The Boldest and Weakest Labor Platforms of the 2020 Democratic Primary." *Vox*, October 29. https://www.vox.com/2019/9/5/20847614/democratic-debate-candidatelabor-platforms.

Card, David, Jörg Heining, and Patrick Kline. 2013. "Workplace Heterogeneity and the Rise of West German Wage Inequality." *Quarterly Journal of Economics* 128 (3): 967–1015.

Chancel, Lucas, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. 2022. *World Inequality Report 2022*. Paris: World Inequality Lab.

Conchon, Aline. 2011. "Board-Level Employee Representation Rights in Europe: Facts and Trends." European Trade Union Institute Report 121.

Dauth, Wolfgang, Sebastian Findeisen, and Jens Suedekum. 2014. "The Rise of the East and the Far East: German Labor Markets and Trade Integration." *Journal of the European Economic Association* 12 (6): 1643–75.

Dauth, Wolfgang, Sebastian Findeisen, Jens Suedekum, and Nicole Woessner. 2021. "The Adjustment of Labor Markets to Robots." *Journal of the European Economic Association* 19 (6): 3104–53.

Der Spiegel. 2022. "Unterschiedliche Löhne bei Tesla in Grünheide sorgen für Unruhe in der Belegschaft." *Der Spiegel*, June 17. https://www.spiegel.de/wirtschaft/unterschiedliche-loehne-beitesla-in-gruenheide-sorgen-fuer-unruhe-in-der-belegschaft-a-3eb597dd-87c54f4e-a4d5-8fb17e7ebde0?sara_ecid=soci_upd_KsBF0AFjflf0DZCxpPYDCQgO1dEMph.

Deutsche Welle. 2022. "Inflation: Germany Seeks to Defuse Social 'Time Bomb.'" *Deutsche Welle*, July 4. https://www. dw.com/en/inflation-germany-seeks-to-defuse-social-time-bomb/a-62353060.

Deutscher Gewerkschaftsbund. 2021. "Mitgliederzahlen 2021." Deutscher Gewerkschaftsbund, Berlin. https://www.dgb.de/uber-uns/dgb-heute/mitgliederzahlen/2020-2029 (accessed March 2, 2022).

DGB. 2021. "Mitgliederzahlen 2021." Deutscher Gewerkschaftsbund. https://www.dgb.de/uber-uns/dgb-heute/ mitgliederzahlen/2020-2029 (accessed March 2, 2022).

DiNardo, John, and David Lee. 2004. "Economic Impacts of New Unionization on Private Sector Employers: 1984–2001." *Quarterly Journal of Economics* 119 (4): 1383–441.

Dustmann, Christian, Bernd Fitzenberger, Uta Schönberg, and Alexandra Spitz-Oener. 2014. "From Sick Man of Europe to Economic Superstar: Germany's Resurgent Economy." *Journal of Economic Perspectives* 28 (1): 167–88.

Dustmann, Christian, Attila Lindner, Uta Schönberg, Matthias Umkehrer, and Philipp Vom Berge. 2022. "Reallocation Effects of the Minimum Wage." *Quarterly Journal of Economics* 137 (1): 267–328.

Dustmann, Christian, Johannes Ludsteck, and Uta Schönberg. 2009. "Revisiting the German Wage Structure." *Quarterly Journal of Economics* 124 (2): 843–81.

Dustmann, Christian, and Uta Schönberg. 2009. "Training and Union Wages." *Review of Economics and Statistics* 91 (2): 363–76.

The Economist. 2017. "The Good and Bad in Germany's Economic Model Are Strongly Linked." July 8. https://www.economist.com/briefing/2017/07/08/the-good-and-bad-ingermanys-economic-model-are-strongly-linked.

The Economist. 2020. "Most of the World Has Yet to Embrace Codetermination." February 1. https://www.economist.com/business/2020/02/01/most-of-the-world-has-yetto-embrace-co-determination.

Ellguth, Peter, and Susanne Kohaut. 2020. "Tarifbindung und betriebliche Interessenvertretung: Aktuelle Ergebnisse aus dem IAB-Betriebspanel 2019." *WSI-Mitteilungen* 73 (4): 278–85.

ETUI. 2021. "National Industrial Relations: Countries." Workers' Participation. https://www.workerparticipation.eu/National-Industrial-Relations/Countries (accessed March 2, 2022).

Eurofound. 1997. "Tripartite Agreement on Employment Alliance for Eastern Germany." Eurofound, June 27. https://www.eurofound.europa.eu/fr/publications/article/1997/tripartite-agreement-on-employment-alliance-for-eastern-germany-0.

Falk, Armin, Ernst Fehr, and Christian Zehnder. 2006. "Fairness Perceptions and Reservation Wages—the Behavioral Effects of Minimum Wage Laws." *Quarterly Journal of Economics* 121 (4): 1347–81.

Farber, Henry, Daniel Herbst, Ilyana Kuziemko, and Suresh Naidu. 2021. "Unions and Inequality over the Twentieth Century: New Evidence from Survey Data." *Quarterly Journal of Economics* 136 (3): 1325–85.

Farber, Henry, and Bruce Western. 2001. "Accounting for the Decline of Unions in the Private Sector, 1973–1998." *Journal of Labor Research* 22 (3): 459–85.

Frandsen, Brigham R. 2021. "The Surprising Impacts of Unionization: Evidence from Matched Employer-Employee Data." *Journal of Labor Economics* 39 (4): 861–94.

Freeman, Richard B., and Edward P. Lazear. 1995. "An Economic Analysis of Works Councils." In *Works Councils: Consultation, Representation, Cooperation in Industrial Relations*, edited by Joel Rogers and Wolfgang Streeck, 27–50. NBER Comparative Labor Markets Series. Chicago: University of Chicago Press.

Gold, Michael, Norbert Kluge, and Aline Conchon. 2010. *'In the Union and on the Board': Experiences of Board-Level Employee Representatives across Europe.* Brussels: European Trade Union Institute.

Goldschmidt, Deborah, and Johannes F. Schmieder. 2017. "The Rise of Domestic Outsourcing and the Evolution of the German Wage Structure." *Quarterly Journal of Economics* 132 (3): 1165–217.

Gürtzgen, Nicole. 2009. "Rent-Sharing and Collective Bargaining Coverage: Evidence from Linked Employer-Employee Data." *Scandinavian Journal of Economics* 111 (2): 323–49.

Handelsblatt. 2022. "Arbeitsminister Heil will Bundesaufträge nur noch an Unternehmen mit Tarifvertrag vergeben." January 1. https://www.handelsblatt.com/politik/deutschland/arbeitsmarkt-arbeitsminister-heil-will-bundesauftraege-nurnoch-an-unternehmen-mit-tarifvertrag-vergeben/27998704.html.

Harlin, Julia D. 1982. "First National Maintenance Corp. V. NLRB: The Supreme Court Narrows Employers' Section 8(A)(5) Duty to Bargain." *Washington and Lee Law Review* 39 (1): 285–302.

Hartung, Benjamin, Philip Jung, and Moritz Kuhn. 2018. "What Hides behind the German labor Market Miracle? Unemployment Insurance Reforms and Labor Market Dynamics." IZA Discussion Paper 12001.

Hassel, Anke. 1999. "The Erosion of the German System of Industrial Relations." *British Journal of Industrial Relations* 37 (3): 483–505.

Haucap, Justus, Uwe Pauly, and Christian Wey. 1999. "The Incentives of Employers' Associations to Raise Rivals' Costs in the Presence of Collective Bargaining." WZB Discussion Paper FS IV 99-6.

Hertel-Fernandez, Alex, William Kimball, and Thomas Kochan. 2022. "What Forms of Representation Do American Workers Want? Implications for Theory, Policy, and Practice." *ILR Review* 75 (2): 267–94.

Hijzen, Alexander, and Pedro S. Martins. 2020. "No Extension Without Representation? Evidence from a Natural Experiment in Collective Bargaining." *IZA Journal of Labor Economics* 9 (5): 1–31.

Hirsch, Boris, and Steffen Mueller. 2020. "Firm Wage Premia, Industrial Relations, and Rent Sharing in Germany." *ILR Review* 73 (5): 1119–46.

IG Metall Bezirk Baden-Württemberg. 2021. "Entgeltrahmen-Tarifvertrag für Beschäftigte der Metall- und Elektroindustrie in Baden-Württemberg." Stuttgart: IG Metall Baden-Württemberg.

International Federation of Robotics (**IFR**)**.** 2017. *World Robotics 2017: Industrial Robots*. Frankfurt: International Federation of Robotics.

ISSP Research Group. 2017. "International Social Survey Programme: Work Orientations IV - ISSP 2015." ZA6770 Data file Version 2.1.0. GESIS Data Archive, Cologne. https://doi.org/10.4232/1.12848.

Jäger, Simon, Benjamin Schoefer, and Jörg Heining. 2021. "Labor in the Boardroom." *Quarterly Journal of Economics* 136 (2): 669–725.

Jäger, Simon, Shakked Noy, and Benjamin Schoefer. 2022a. "Codetermination and Power in the Workplace." *ILR Review* 75 (4): 857–90.

Jäger, Simon, Shakked Noy, and Benjamin Schoefer. 2022b. "What Does Codetermination Do?" *ILR Review* 75 (4): 857–90.

Jäger, Simon, Shakked Noy, and Benjamin Schoefer. 2022c. "Replication data for: The German Model of Industrial Relations: Balancing Flexibility and Collective Action." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E176722V1.

Jirjahn, Uwe, and Stephen C. Smith. 2018. "Nonunion Employee Representation: Theory and the German Experience with Mandated Works Councils." *Annals of Public and Cooperative Economics* 89 (1): 201–34.

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

Kay, Grace. 2022. "Germany's biggest auto union questions Elon Musk's authority to give a return-to-office ultimatum: 'An employer cannot dictate the rules just as he likes'." *Business Insider*, June 9. https://www.businessinsider.in/thelife/news/germanys-biggest-auto-union-questions-elon-musks-authority-to-give-a-return-to-office-ultimatum-an-employer-cannot-dictate-the-rules-just-as-he-likes/articleshow/92092087.cms.

Kell, Alexander. 2022. "Germany's Union Head Warns of Collapse of Entire Industries." *Bloomberg*, July 3. https://www.bloomberg.com/news/articles/2022-07-03/germany-s-unionhead-warns-of-collapse-of-entire-industries.

Krebs, Tom, and Martin Scheffel. 2013. "Macroeconomic Evaluation of Labor Market Reform in Germany." *IMF Economic Review* 61 (4): 664–701.

Kügler, Alice, Uta Schönberg, and Ragnhild Schreiner. 2018. "Productivity Growth, Wage Growth and Unions." In *Price and Wage-Setting in Advanced Economies*, 215–247. Frankfurt: European Central Bank.

Lee, David S., and Alexandre Mas. 2012. "Long-Run Impacts of Unions on Firms: New Evidence from

Financial Markets, 1961–1999." *Quarterly Journal of Economics* 127 (1): 333–78.
Lepinski, Franz. 1959. "The German Trade Union Movement." *International Labour Review* 79 (1).
Lesch, Hagen. 2009. "Erfassung und Entwicklung von Streiks in OECD-Ländern." Köln: Institut der deutschen Wirtschaft.
Matthews, Dylan. 2019. "Unions for All": The New Plan to Save the American Labor Movement." *Vox*, September 2. https://www.vox.com/policy-and-politics/2019/9/2/20838782/unions-forall-seiu-sectoral-bargaining-labor-unions.
Moene, Karl Ove, Michael Wallerstein, and Michael Hoel. 1992. "Bargaining Structure and Economic Performance." Unpublished.
Müller, Torsten, and Thorsten Schulten. 2019. "Germany: Parallel Universes of Collective Bargaining." In *Collective Bargaining in Europe: Towards an Endgame*, edited by Torsten *Müller, Kurt* Vandaele, and Jeremy Waddington, 239–65. Brussels: European Trade Union Institute.
Oberfichtner, Michael, and Claus Schnabel. 2018. "The German Model of Industrial Relations: (Where) Does It Still Exist?" *Jahrbücher* für *Nationalökonomie und Statistik* 239 (1): 5–37.
Ochel, Wolfgang. 2005. "Decentralising Wage Bargaining in Germany—A Way to Increase Employment?" CESifo Working Paper 1069.
OECD. 2019. *Negotiating Our Way Up: Collective Bargaining in a Changing World of Work*. Paris: OECD. https://doi.org/10.1787/1fd2da34-en.
OECD. 2021. "OECD/AIAS ICTWSS Database." 'https://www.oecd.org/employment/ictwss-database.htm (accessed April-04-2022).
OECD. 2022a. "Average wages (indicator)." OECD iLibrary. https://doi.org/10.1787/cc3e1387-en (accessed April 4, 2022).
OECD. 2022b. "Employee compensation by activity (indicator)." OECD iLibrary. https://doi.org/10.1787/ 7af78603-en (accessed April 4, 2022).
OECD. 2022c. "GDP per hour worked (indicator)." OECD iLibrary. https://doi.org/10.1787/1439e590-en (accessed April 4, 2022).
OECD. 2022d. "Hours worked (indicator)." OECD iLibrary. https://doi.org/10.1787/47be1c78-en (accessed April 4, 2022).
OECD. 2022e. "Labour force participation rate (indicator)." OECD iLibrary. https://doi.org/10.1787/8a801325-en (accessed April 4, 2022).
OECD. 2022f. "Unemployment rate (indicator)." OECD iLibrary. https://doi.org/10.1787/52570002-en (accessed April 4, 2022).
OECD. 2022g. "Value added by activity (indicator)." OECD iLibrary. https://doi.org/10.1787/a8b2bd2ben (accessed April 4, 2022).
OECD. 2022h. "Wage levels (indicator)." OECD iLibrary. https://doi.org/10.1787/0a1c27bc-en (accessed April 4, 2022).
Paster, Thomas. 2012. "Do German Employers Support Board-Level Codetermination? The Paradox of Individual Support and Collective Opposition." *Socio-Economic Review* 10 (3): 471–95.
Price, Brendan. 2018. "The Duration and Wage Effects of Long-term Unemployment Benefits: Evidence from Germany's Hartz IV Reform." Unpublished.
Raess, Damian. 2014. "Export Dependence and Institutional Change in Wage Bargaining in Germany." *International Studies Quarterly* 58 (2): 282–94.
Raymunt, Monica. 2022. "Tesla's German Hiring Hitting Roadblocks Over Wages, Union Says." *Bloomberg*, June 20. https://www.bloomberg.com/news/articles/2022-06-20/tesla-s-germanhiring-hitting-roadblocks-over-wages-union-says.
Rehder, Britta. 2003. *Betriebliche Bündnisse für Arbeit in Deutschland: Mitbestimmung und Flächentarif im Wandel*. Frankfurt: Campus Verlag.
Schnabel, Claus. 2020. "Betriebliche Mitbestimmung in Deutschland: Verbreitung, Auswirkungen und Implikationen." *Perspektiven der Wirtschaftspolitik* 21 (4): 361–78.
Schulten, Thorsten, Götz Bauer, Marion Frömming, Rosemarie Pulfrich, Andrea Taube, Monika Wollensack, and Jasmina Ziouziou. 2021. *2021 Tarifpolitik: Statistisches Taschenbuch*. Düsseldorf: Wirtschafts- und Sozialwissenschaftliches Institut.
Schulten, Thorsten, and Reinhard Bispinck. 2018. "Varieties of Decentralisation in German Collective Bargaining." In *Multi-Employer Bargaining Under Pressure: Decentralisation Trends in Five European Countries*, edited by Salvo Leonardi and Roberto Pedersini, 105–49. Brussels: European Trade Union Institute.
Seifert, Hartmut, and Heiko Massa-Wirth. 2005. "Pacts for Employment and Competitiveness in Germany." *Industrial Relations Journal* 36 (3): 217–40.
Sick, Sebastian. 2020. "Erosion als Herausforderung für die Unternehmensmitbestimmung."

Mitbestimmung-Portal, May 1. https://www.mitbestimmung.de/html/erosion-als-herausforderung-furdie-14188.html.

Silvia, Stephen J. 1997. "German Unification and Emerging Divisions within German Employers' Associations: Cause or Catalyst?" *Comparative Politics* 29 (2): 187–208.

Silvia, Stephen J. 2013. *Holding the Shop Together: German Industrial Relations in the Postwar Era*. Ithaca: Cornell University Press.

Silvia, Stephen J., and Wolfgang Schroeder. 2007. "Why Are German Employers Associations Declining? Arguments and Evidence." *Comparative Political Studies* 40 (12): 1433–59.

Soziale Politik für Dich (SPD). 2017. *Zeit für mehr Gerechtigkeit: Unser Regierungsprogramm für Deutschland*. Berlin: Soziale Politik für Dich.

Soziale Politik für Dich (SPD). 2021. *The SPD's Programme for the Future*. Berlin: Soziale Politik für Dich.

Sozialdemokratische Partei Deutschlands (SPD), Bündnis 90 / Die Grünen, and Freie Demokraten Partei (FDP). 2021. *Mehr Fortschritt Wagen: Bundnis fur Freiheit, Gerechtigkeit und Nachhaltigkeit*. Berlin: SPD, Gruene, and FDP.

Stansbury, Anna, and Lawrence H. Summers. 2020. "The Declining Worker Power Hypothesis: An Explanation for the Recent Evolution of the American Economy." NBER Working Paper 27193.

Stettes, Oliver. 2007. "Die Arbeitnehmermitbestimmung im Aufsichtsrat: Ergebnisse einer Unternehmensbefragung." *IW-Trends-Vierteljahresschrift zur empirischen Wirtschaftsforschung* 34 (1): 17–30.

Strine Jr., Leo, Aneil Kovvali, and Oluwatomi Williams. 2021. "Lifting Labor's Voice: A Principled Path Toward Greater Worker Voice and Power within American Corporate Governance." Unpublished.

Thelen, Kathleen A. 1991. *Union of Parts: Labor Politics in Postwar Germany*. Ithaca: Cornell University Press.

Thelen, Kathleen A. 2014. *Varieties of Liberalization and the New Politics of Social Solidarity*. Cambridge, UK: Cambridge University Press.

Umkehrer, Matthias. 2017. *Combining the Waves of the IAB Establishment Panel: A Do-File for the Basic Data Preparation of a Panel Data Set in Stata*. Nuremberg: German Federal Employment Agency, Institute for Employment Research.

Visser, Jelle. 2021. *OECD/AIAS ICTWSS Database: Detailed Note on Definitions, Measurement, and Sources*. Paris: OECD.

Wang, Sean, and Samuel Young. 2022. "Unionization, Employer Opposition, and Establishment Closure." Unpublished.

Weil, David. 2014. *The Fissured Workplace: Why Work Became So Bad for So Many and What Can Be Done to Improve It*. Cambridge, MA: Harvard University Press.

Western, Bruce, and Jake Rosenfeld. 2011. "Unions, Norms, and the Rise in U.S. Wage Inequality." *American Sociological Review* 76 (4): 513–37.

Winkler, Heinrich August. 1998. *Weimar 1918-1933: Die Geschichte der ersten deutschen Demokratie*. Munich: C.H. Beck.

Wochenblatt-Online. 2020. "RAFI und IG Metall einigen sich auf neuen Anerkennungstarifvertrag (ATV)." *Wochenblatt-Online*, October 21. https://wochenblatt-online.de/rafi-und-ig-metall-einigen-sichauf-neuen-anerkennungstarifvertrag-atv/.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 81–102

Danish Flexicurity: Rights and Duties

Claus Thustrup Kreiner and Michael Svarer

Denmark is a small country with 5.8 million inhabitants that achieves a high income per capita in combination with low inequality and comprehensive social insurance. Table 1 provides statistics on happiness and key indicators on economic performance and public policy for Denmark and the United States, including how these countries rank among OECD countries, from #1 ("best") to #36 ("worst"), based on these indicators. Denmark ranks first in this comparison group in happiness (row 1 of Table 1). Labor market performance in Denmark as measured by employment (row 7), long-term unemployment (row 8), and labor market turnover (row 9) is comparable to the United States, but Denmark achieves this in combination with a generous unemployment compensation scheme with long duration of unemployment insurance benefits (row 17) and high compensation rates for people with low earnings (rows 18–19).

The Danish labor market model has come to be known as "flexicurity." A stated strategy underlying this approach is the so-called "*right and duty*" principle (in Danish, "ret og pligt"). Unemployed individuals have a *right* to receive income

■ *Claus Thustrup Kreiner is Professor of Economics and Director of Center for Economic Behavior and Inequality (CEBI), University of Copenhagen, Copenhagen, Denmark. He is also Area Director of Public Economics in the CESifo network, Münich, Germany, and Research Fellow, Centre for Economic Policy Research, London, United Kingdom. Michael Svarer is Professor of Economics, Aarhus University, Aarhus, Denmark, and Research Fellow, IZA Institute of Labor Economics, Bonn, Germany. Both authors have served on commissions concerning taxation and labor market policy and have co-chaired The Danish Economic Councils. Their emails are ctk@econ.ku.dk and msvarer@econ.au.dk.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.81.

Table 1
Performance and Policy Parameters of Denmark and the United States

	Value		*OECD rank (1–36)*	
	Denmark	*United States*	*Denmark*	*United States*
Performance				
1. Subjective happiness (0–10 scale) (+)	7.6	6.9	1	15
2. Economic freedom (index 0–100) (+)	76.7	76.8	10	8
3. Confidence in government (%) (+)	63	31	6	30
4. Income per capita (thousands of US$) (+)	62	66	7	5
5. Inequality: Gini (%) (–)	26	39	6	33
6. Low pay incidence (%) (–)	8	24	3	36
7. Employment rate (%) (+)	75	71	12	20
8. Share long-term unemployed (%) (–)	17	13	9	8
9. Labor market turnover (%) (+)	22	20	5	7
Policy				
10. Tax burden (% of GDP) (–)	46	24	36	5
11. Social spending (% of GDP) (+)	28	19	4	20
12. Spending, passive LMP (% of GDP) (+)	1.1	0.2	10	34
13. Spending, active LMP (% of GDP) (+)	2.0	0.1	1	33
14. Public share of education spending (%) (+)	98	68	1	32
15. Union density (% of workforce) (+)	66	10	3	31
16. Employment protection (index 0–6) (+)	1.8	1.3	29	36
17. Unemployment insurance benefit duration (months) (+)	24	6	5	28
18. Net replacement rate, 3 months (%) (+)	83	57	4	27
19. Net replacement rate, 3rd year (%) (+)	67	8	1	33

Sources: OECD (2017a, b, 2018a, b, c, d, 2019a, b, c, d, e, f, g, h, i, j, k, 2020a, b) and Miller, Kim, and Roberts (2019). Details in online Data Repository (Kreiner and Svarer 2022).
Notes: The first parenthesis to the right of the row title shows the unit of measurement of the indicator. In the second parenthesis, a (+) indicates that countries with higher values of the indicator in columns 1–2 are ranked better when computing the rankings in columns 3–4, while a (−) indicates that countries with higher values are ranked worse. Rankings are among the 36 OECD countries based on the indicators, where "1" is best and "36" is worst. If Denmark or the United States have the exact same value as another country, then as a convention, we give Denmark/US the best rank number. Data is from 2019 or latest available year. Income per capita corresponds to GNI. Low-paid workers denotes the percentage of full-time workers earning less than two-thirds of gross median earnings. Share long-term unemployed is the percentage of unemployed individuals who have been unemployed for longer than 12 months. Labor market turnover is the difference between the hiring rate and the net employment change. LMP denotes Labor Market Policies. Employment protection is an OECD average score of four broad indicators of worker protection. Net replacement rates are for a single person with no children, earning 67 percent of the average wage level prior to unemployment.

support and to receive public assistance in getting back into work. But it is also their *duty* to search actively for jobs, to take on appropriate work, and to participate in active labor market policies. Correspondingly, society has a *right* to make demands of recipients of income support, but also a *duty* to help improve their job prospects.

In this essay, we begin with a description of flexicurity and compare Danish labor market policy and performance to the United States and other OECD countries. Some labor markets, in particular in the Nordic countries, share key similarities, but none of them have all the characteristics of the flexicurity model. We

then look more closely at the history and formation of Danish flexicurity policy and labor market development, and in particular to extensive reforms that changed key elements of the program in the early 1990s. Key to the Danish flexicurity model is massive spending in "active labor market programs," with compulsory participation for recipients of unemployment compensation. We review the theoretical foundation for this policy as well as the microeconometric evidence on its effect on the employment prospects of the unemployed. We also discuss the aptness of flexicurity policy to meet challenges from globalization, automation, and immigration. The last section concludes and discusses some issues that the United States (or other countries) would face in adopting a flexicurity policy.

The Danish Labor Market

Collective Bargaining

The Danish labor market model is the product of a long tradition of organized bargaining between workers and employers.[1] In Denmark, collective negotiations between unions and employer organizations dating back to the so-called September Agreement of 1899 have decided key labor market conditions, like hourly wages and hours worked. The original agreement followed a labor dispute of more than 100 days involving strikes and lockdowns (for discussion, see Høgedahl 2020). At one point during the dispute, more than half of the organized labor force was locked out. The dispute ended with an agreement that the employers accept the workers' right to organize and the unions accept the employers' right to manage.

In its current form, the bargaining follows specified rules and a so-called "conciliation institution" helps in solving disagreements. The government is typically a passive partner in these negotiations, but if the parties cannot reach an agreement, the government can intervene and even dictate agreements. Lockouts and strikes can occur during the formal negotiation periods but are illegal between these periods.

Denmark has never had a statutory minimum wage. Basic wage levels are typically negotiated by trade unions and employer organizations at the sector level, and the final wage-setting is often determined in local negotiations at the firm level (for more details on the development of the wage negotiations in the Danish labor market, see Dahl, le Maire, and Munch 2013). In wage negotiations, unions are represented by larger trade union confederations. The largest is the Danish Trade Union Confederation (FH). It represents 64 different member organizations that each represent one or more occupations. FH bargains at the national level with the Confederation of Danish Employers (DA). The DA/FH area covers around half of the private labor market and has typically negotiated the first agreement, which then becomes a benchmark for the remaining agreements in the labor market, including for the public sector.

[1] More details on the Danish labor market can be found in Hansen and Tranæs (1999), Andersen and Svarer (2007), and Andersen (2019).

Unions play a large role in Danish society. Besides being a main part of the wage negotiations, they also take part in political processes on labor market policy. The so-called "triparty agreements" between the government, the employer associations, and the unions are the customary way to make decisions regarding labor market policies, educational policies, work safety, and other issues.

A precondition for such a system is a strong collective bargaining system. Denmark has a union density of around 66 percent—among the OECD countries with the highest union density (row 15). In contrast, the United States has a union density of only 10 percent. Similarly, the coverage of the collective bargaining systems (that is, the share of workers whose wages are determined by collective bargaining even if they are not personally members of a union) is 84 percent in Denmark and 12 percent in the United States (OECD 2017c).

Flexicurity

"Flexicurity" describes the Danish labor market policy that combines flexible hiring and firing rules for firms with high income security for workers. Making it easy to hire and fire workers allows each firm to adjust worker input in production and ensures high production efficiency and economic growth. Job security is low, but this is acceptable to workers and unions due to generous income compensation when unlucky workers are hit by temporary job losses, combined with an active labor market policy that helps such workers back into employment.

One measure of flexibility in hiring and firing decisions of firms is the OECD Employment Protection Index (row 16). This flexibility is similar in both the United States and Denmark: that is, both countries provide little job security and make it easy for employers to adjust their labor force. As a result, labor market turnover rates of the United States and Denmark are similar and at a high level compared to other countries (row 9). The high freedom of firms to adjust labor input aligns with the more general Index of Economic Freedom (row 2), where Denmark and the United States are also aligned.

But while labor market flexibility is very similar in Denmark and the United States, the income security provided for unemployed workers is very different. Denmark is ranked near the top of high-income countries, both in terms of the maximum duration of unemployment benefits of two years and in terms of unemployment compensation, where the net replacement rate is 83 percent after three months of unemployment for people in the lower part of the wage distribution (rows 17–18). The United States is at the other end of the spectrum, with a maximum unemployment duration of six months under normal business cycle conditions, and with a net replacement rate (for a low-income single childless person) of 57 percent after three months of unemployment. In Denmark, unemployment insurance is partly paid from employer contributions to a fund, but also heavily subsidized by the government.[2]

[2] The unemployment benefit scheme in the United States includes the likelihood of extended unemployment insurance during recessions, which is not reflected in the table. Also, the net replacement rate varies

The difference in income security becomes more striking in the third year of unemployment, at which point unemployment benefits are exhausted in both countries. For a low-income single person without children, it is possible to get means-tested benefits corresponding to a net replacement rate of up to 67 percent in Denmark, compared with 8 percent in the United States (row 19).[3]

The more generous benefit system in Denmark is reflected in the total spending on "passive" labor market policies—that is, policies like unemployment insurance that just provide payments to individuals—which is above 1 percent of GDP compared to 0.15 percent in the United States (row 12). A standard worry of economists is that this high generosity might dampen incentives to work and reduce employment. However, this concern is not reflected in Denmark's labor market. The Danish employment rate is higher than in the United States, and the two countries are ranked similarly to each other when it comes to the incidence of long-term unemployment (rows 7–8).[4]

Active Labor Market Programs

Unemployed individuals in Denmark are required to participate in "active" labor market programs, which is a central component of the Danish flexicurity model. These programs provide job-search assistance, work practice, and retraining in exchange for receiving unemployment benefits. Unemployment benefits can be sanctioned if an unemployed fails to comply with the requirements. In 2021, around 12 percent of the unemployed were sanctioned at some point of their unemployment spell.[5] The total costs of active labor market programs are close to 2 percent of GDP per year and make Denmark, by a wide margin, the OECD country that spends most on active labor market policy (row 13). The United States, at the other end of the spectrum, allocates 0.1 percent of GDP to active labor market measures.

The intensity and duration of active labor market policies increases during a period of unemployment. The unemployed are matched to a caseworker. In the early stages of an unemployment spell, they meet regularly and the caseworker monitors job search activities and guides the job search process. The first meeting occurs within one month of unemployment. If deemed necessary, an unemployed person can participate in short job search courses. If the caseworker assesses that an unemployed

---

with previous income and family characteristics. Table 1 is based on single individuals earning 67 percent of the average wage. The difference between Denmark and the United States is smaller when looking at families with children, but the replacement rate is in all cases larger in Denmark. Danish workers are better insured against job loss, but it is worth noting that for a given loss in disposable income, the drop in consumption is similar across Denmark and the United States (Andersen et al. 2021).

[3] In Denmark, this includes the guaranteed minimum income benefit and housing benefit programs, while for the United States it includes the Supplemental Nutrition Assistance Program. For more details on the mean-tested benefit programs and the computation of the replacement rates, see OECD (2020c, 2020d, 2020e).

[4] The higher employment rate in Denmark does not imply that overall labor input in Denmark is higher than in the United States. Hours worked per person is considerably lower in Denmark and the other Nordic countries compared to the United States, because of both fewer work weeks and lower weekly work hours (Bick, Brüggemann, and Fuchs-Schündeln 2019).

[5] Reported by The Danish Agency for Labor Market and Recruitment of the Ministry of Employment at http://jobindsats.dk.

person needs educational requalification or closer contact to the labor market to increase job chances, it is possible to engage in four-week work practice jobs at public or private firms or to participate in short employment-focused educational programs. If these short-term measures are insufficient to bring an unemployed person back into employment, it is possible to have longer subsidized employment periods of up to four months duration in either private or public companies or to engage in long-term educational programs. These activities typically start after six months of unemployment.

Strong unions and generous unemployment benefits affect the wage distribution, in particular by ensuring a high effective minimum wage floor. In Denmark, only 8 percent of employees work in full-time jobs that pay less than two-thirds of the gross median earnings, whereas in the United States it is close to one-quarter (row 6). Again, Denmark and the United States are in the opposite end of the rank distribution among OECD countries.

A high minimum wage floor risks excluding low-productivity individuals from entering the job market. Some people who are eager to work might have productivity levels below the required threshold. One purpose of the active labor market policy—and, more generally, the education system—is to ensure that nobody falls below the minimum-productivity threshold. In Denmark, education at all levels is provided free-of-charge by the public sector, with almost no role played by private institutions: overall, the government share of total education expenditures is 98 percent (row 14). This includes substantial resources devoted to adult vocational training of employed workers at off-the-job training sites. Denmark is the only OECD country where the public sector provides and finances this type of vocational training (Humlum and Munch 2019). In addition, adult students receive student allowances and access to cheap government loans.

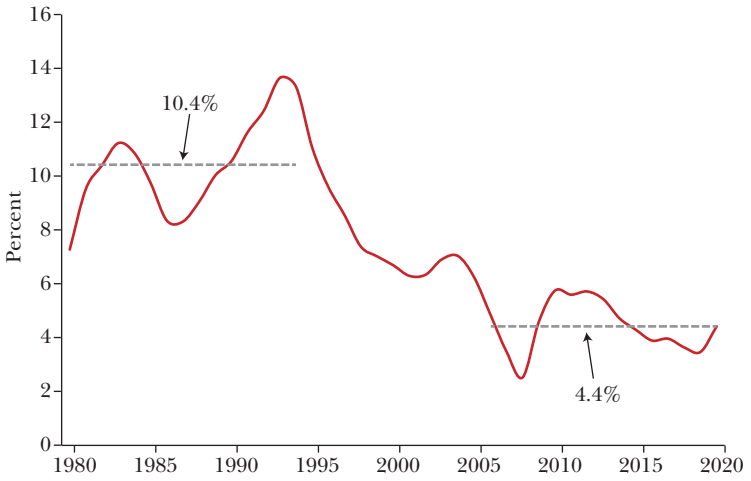
The History and Evolution of the Flexicurity Model

Denmark has a long history of combining a high degree of flexibility in hiring and firing decisions of firms with a high level of income security. However, the flexicurity model underwent major changes in the early 1990s. Here we discuss the shift that occurred.

Failure of the Old Flexicurity Regime

The older flexicurity model had even longer maximum duration of unemployment insurance benefits than the two years today. In practice, the duration was close to infinity, because participation in active labor market programs at the end of the statutory 2.5-year duration of unemployment insurance benefits was sufficient to qualify for a new 2.5-year period. Unemployed workers were offered job training and education in active labor market programs, but with a much lower intensity and with voluntary participation—which in practice started close to expiration of unemployment benefits.

Figure 1
Unemployment Rate in Denmark



Source: Statistics Denmark (1980–2020).
Notes: The graph plots the share of people in the labor force who are unemployed. It is based on administrative records of people who are registered as unemployed and includes people who participate in active labor market programs. The gray dashed lines are averages for 1980–1994 and 2006–2020.

The pre-1990 flexicurity model failed to combine high income security with low unemployment. Figure 1 plots the unemployment rate in Denmark over the last four decades, based on administrative records of people who are registered as unemployed and including people who participate in active labor market programs. After the oil price shocks and macroeconomic disruption of the 1970s, the share of unemployed people in the labor force reached 10 percent in the early 1980s. This was expected to be temporary. But while the favorable business cycles in the mid-1980s brought unemployment down to 8 percent in 1986–1987, it also led to significant wage rises. The nominal hourly wage rate in the industry sector grew annually by 7 percent in 1986–1987, corresponding to an annual real wage growth of 4 percent, and the total wage share out of gross factor income increased from 54 percent to 58 percent from 1984 to 1987 (for details, see Kreiner and Svarer 2022; Danish Economic Council 1995). Afterwards, unemployment climbed to 14 percent in 1993.

Over the 15-year period of the old flexicurity regime, from 1980 to 1994, unemployment fluctuates around an average, long-run rate of more than 10 percent, as illustrated by the horizontal, dashed line in the left part of Figure 1. In addition, survey evidence in Pedersen and Smith (1995) shows that 40 percent of the unemployed recipients of unemployment insurance in the early 1990s did not fulfill standard international criteria for being unemployed by being ready to take up relevant work and actively searching for a job (ILO 2019). Thus, a large share of the recipients of unemployment benefits did not seem to be involuntary unemployed.

Toward a New Flexicurity Regime

The poor labor market performance in the old flexicurity regime triggered major adjustments starting in the early 1990s. The flexibility in hiring and firing of firms was unchanged, but income security decreased. The maximum duration of unemployment insurance was reduced to four years at the turn of the century and was then reduced further to its current length of two years. However, even after exhaustion of unemployment insurance benefits, unemployed workers are still eligible for means-tested social assistance at a relatively high level. As noted earlier, a low-paid person without children can still receive up to two-thirds of previous income (row 19 in Table 1). The replacement rate is reduced significantly for a few targeted groups, most notably young workers under age 25. These targeted reductions in income security did appear to increase employment to some extent (for example, Jonassen 2013; Danish Economic Council 2014; Hermansen 2015).

However, by far the biggest change was in the area of active labor market policy. A major labor market reform in 1994 introduced the key principle of "rights and duties" into the active labor market policy. Recipients of unemployment insurance, as well as people receiving social assistance, are now required to apply for jobs, to participate in active labor market policies, and to accept job offers fitting their profiles. Failure to comply is met with benefit sanctions (Svarer 2011). In addition, the active labor market policy intensified by offering and requiring participation frequently in active labor market policies. In 1993–1994, before the reforms, one out of six unemployed individuals participated in a program during the year; in comparison, during the last decade more than half of the unemployed participated annually in some type of activation program (Ministry of Employment 1993–2019; for more details, see Kreiner and Svarer 2022).

This "workfare" element can increase the willingness to work of unemployed individuals and moderate wage claims of workers and unions because of a worsening of their threat point/outside option in the wage negotiations. At the same time, the programs can increase productivity of workers and reduce information frictions—and thereby increase employment. In Figure 1, note that the revised flexicurity regime was followed by a descent of the unemployment rate over the next 15 years to a much lower long-run level. During the last 15 years, the unemployment rate fluctuates around a long-run level of 4.4 percent. We attribute this major improvement in labor market performance mainly to the intensified Danish active labor market policy, alongside the changes in the unemployment insurance scheme. In the next section, we discuss the theoretical foundation for introducing workfare in the active labor market policy and review microeconometric studies on the employment effects of the Danish active labor market policy.[6]

[6]One may ask about the role of earned income tax credits (EITCs) to boost employment in this case. An EITC was implemented in Denmark much later (2004) than in the United States and, therefore, cannot explain the big drop in unemployment, which took place earlier. Moreover, participation tax rates continue to be high in Denmark because of the high out-of-work benefits (Kleven and Kreiner 2005; Immervoll et al. 2007).

Impact of the Collective Bargaining System?

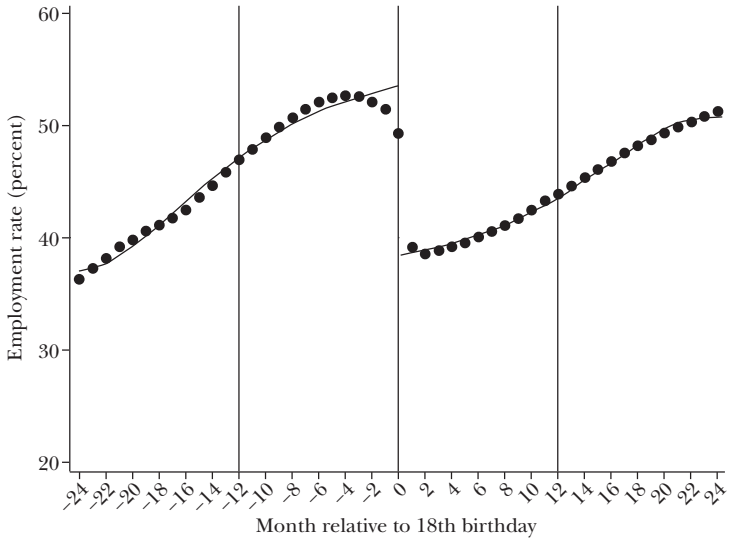
A theoretical hypothesis is that a collective bargaining system can achieve both high wages and high employment, with lower-skilled workers being paid more than their productivities. Can this explain the successful Danish labor market performance? In the efficient bargaining model of McDonald and Solow (1981), such an outcome is possible because both wages and employment are the subject of negotiation between unions and employer organizations. However, in the main agreement between Danish unions and employer organizations dating back to 1899, it is stated explicitly that firms have the right to manage—that is, the right to decide on hiring and firing of workers. In this case, where bargaining is only over wages and firms decide employment, theory suggests that firms do not keep workers with productivities below the going wage.

It could still be the case that workers and firms agree explicitly or implicitly on wage compression, where firms combine more-skilled people paid below their productivity level and less-skilled people paid above their productivity level. In this case, employment of less-skilled people is not on the labor demand curve, as firms pay this group more than their productivity level. Empirical evidence for young people, who are low-skilled and earn low wages, indicates that this does not take place in practice. Kreiner, Reck, and Skov (2020) use population records on wages and employment at the monthly frequency to study what happens when young workers turn 18 years old and become eligible for the significantly higher negotiated minimum wages that apply for adults. In the agreements, the basic minimum hourly wage rate of a young adult who is 18 years or older was around $15 in 2016. It is considerably lower when younger. On average, the observed wage rate jumps up by 40 percent at age 18 (computed using the midpoint method), and this jump is of a similar size as the jump in the agreed minimum wage levels. Thus, minimum wages appear to be binding.

Figure 2 shows how the minimum wage hike at age 18 affects employment. The figure shows monthly employment rates for people at age 16–20. At the age discontinuity of the minimum wage, employment of young workers drops by 15 percentage points. This implies that one-third of the employed lose their job when they turn 18 years old. The graph also shows that it takes two additional years (age 20) before employment is back to the level before the wage hike. The quick employment adjustment of firms at the wage hike strongly suggests that employment is on the labor demand curve and, importantly, that firms in Denmark do not keep low-skilled workers if wage costs are above their productivity levels.

Danish wage setting became more decentralized during the 1990s, with a large part of wages being determined in bargaining at the firm level (Boeri, Brugiavini, and Calmfors 2001). This greater flexibility in the wage determination also led to more wage dispersion (Dahl, le Maire, and Munch 2013). This shift may have contributed to the rise in employment by making wages in the lower part of the wage distribution more aligned with productivity levels. On the other hand, minimum wages and many key labor market conditions continue to be negotiated at the sector level. The organizational changes seem too small to fully explain the big decline in

Figure 2
Employment Rate around Workers' 18th Birthday



Source: Kreiner et al. (2020).
Note: The figure depicts employment rates by age, in months, for two years before and after individuals turn 18 years old. It is based on monthly payroll records for the Danish population. The figure replicates Figure 1.B in Kreiner et al. (2020), which describes the data and the estimation of the fitted line and the percentage drop in employment at age 18. The graph shows that employment drops by 15 percentage points, or 33 percent, when people turn 18, where the wage rate jumps up by 40 percent. The percentage changes are computed using the midpoint method.

unemployment. Again, we see the major change in the flexicurity policy as the likely most important driver of the long-run development in unemployment.

To sum up: The highly organized labor market ensures that low-skilled and vulnerable workers are not exploited and receive decent wages. Together with a generous unemployment compensation scheme, this creates high income security, while the flexibility in hiring and firing decisions of firms supports a high labor demand. However, to keep employment at high levels, it is important to also spend large resources on active labor market policy and to include both carrots and sticks.

Active Labor Market Programs in Theory and Practice

What are the potential benefits of a "workfare" policy in which recipients of unemployment benefits must spend time in certain government-organized active labor market policies? In empirical terms, does Denmark's high spending on active labor market policy significantly enhance labor market prospects of the participants?

Workfare Can Mitigate Adverse Selection

To isolate the role of a workfare component in active labor market policy we may ask: Can it be socially optimal to require workfare activities in exchange for unemployment benefits if the activities themselves are unproductive, like the equivalent of digging holes and re-filling them? The answer is yes (Hansen and Tranæs 1999; Kreiner and Tranæs 2005).[7]

To see why, recall that under the old flexicurity regime in Denmark without workfare, 40 percent of unemployment insurance recipients did not fulfill standard criteria for being involuntarily unemployed. In this case (of adverse selection), workfare can be used as a screening device to prevent people with more taste for leisure from claiming high unemployment insurance benefits intended for involuntary unemployed individuals. This is illustrated in Figure 3. It shows two examples of labor supply decisions for two individuals X and Y. In both diagrams, the budget line illustrates how extra hours of work h increases disposable income y, with the slope of the budget line given by the net-of-tax wage rate. Utility is increasing when moving northwest in the diagrams corresponding to getting more income and more leisure. Indifference curves I_2^X and I_1^X illustrate preferences of type X, while the indifference curve I^Y illustrates an indifference curve of type Y, which is less eager to work than type X. Type X always prefers point A and working h^* hours.

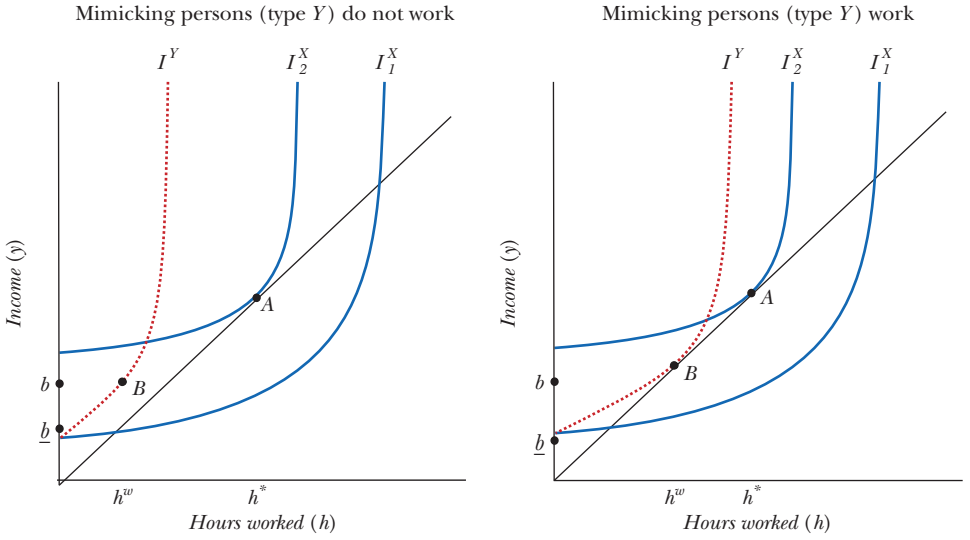
Consider the case where it is possible for those who are not working to receive social assistance $\underline{b}$ but no unemployment benefits b. In the left panel, type Y prefers to receive social assistance $\underline{b}$ instead of working. The policymaker would like to offer unemployment benefits b to type X individuals who cannot find a job and are involuntary unemployed. However, the policymaker cannot distinguish between an involuntary unemployed type-X person and a type-Y person who currently does not wish to work at the going wage. Offering unemployment insurance benefits to involuntary unemployed individuals of b is costly because type Y individuals can also claim these benefits—and will do so if b is higher than $\underline{b}$.

But if receiving b is made conditional on spending h^w hours on workfare activities (point B in the figure), then it is not attractive for type Y, who in this case prefers to receive $\underline{b}$, which is not conditioned on workfare (notice that point B is on the indifference curve for type Y, while $\underline{b}$ is just above this indifference curve).

Unemployed type-X individuals will claim the combination of benefits and workfare offered in point B, and only if they cannot find a job. This gives them a strictly higher utility level than I_1^X, corresponding to social assistance $\underline{b}$. Therefore, the policymaker can make a Pareto improvement by offering point B, compared to a situation with only the social assistance level $\underline{b}$ (for a formal proof, see Kreiner and Tranæs 2005).

[7] Here, we study the use of workfare in the context of active labor market policy and involuntary unemployment, and show it can be Pareto-optimal to use workfare. Another strand of literature asks whether it is socially optimal to require unproductive workfare activities of low-skilled people as part of redistribution policy (Besley and Coate 1992, 1995). In this context, the "screening problem" is different and it is typically not Pareto-optimal to use workfare.

Figure 3
Optimal Use of Workfare in Unemployment Compensation Schemes



Source: Authors' illustrations.
Note: The graphs plot income y by hours worked h of two individuals (X and Y). They have the same budget line, but different preferences for work illustrated by their indifference curves. Type X is most eager to work and chooses point A in both panels if working. However, type X may be temporary jobless. In this case, in the left panel, offering the unemployment insurance benefit level b to involuntary unemployed type X persons is too costly because individuals outside the labor market (type Y) who normally receive the social assistance benefit level b will also claim the high benefit level $\underline{b}$. Requiring recipients of b to spend h^w hours in workfare activities avoids this mimicking and targets the high benefit level b to unemployed type X persons. This increases their utility while keeping the benefit and utility level of type Y persons unchanged. The right panel illustrates a similar situation where a mimicking type Y person will stop working at point B if it is possible to claim the high benefit level b. By requiring that benefit recipients spend h^w hours in workfare activities, it is possible to target the high benefit level b to unemployed type X persons and increase their utility.

The right panel illustrates another type of case where it can also be socially optimal to use workfare. In this case, the indifference curve I^Y is such that type Y prefers to work at point B compared to receiving social assistance $\underline{b}$ (note that the indifference curve is tangent to the budget line and that the indifference curve is above $\underline{b}$). Introducing unemployment benefits b without workfare to involuntary unemployed is again costly, this time because type Y will stop working and claim benefits. However, if such benefits are combined with h^w hours of workfare (point B) then this is not more attractive than working for type Y. Type X strictly prefers unemployment benefits combined with workfare at point B compared to receiving social assistance $\underline{b}$. Therefore, the policymaker can make a Pareto improvement by offering point B compared to a situation with only the social assistance level $\underline{b}$. Thus, also in

this case, workfare can be an attractive tool for policymakers who wish to offer high unemployment compensation for the involuntary unemployed.

To conclude, requiring participation in active labor market programs may work as a "screening device" that prevents some people from becoming voluntary unemployed and receiving unemployment insurance benefits.

Workfare Can Mitigate Moral Hazard and Enhance Competencies

In addition to these results on adverse selection, complementary research shows in an equilibrium search-setting that workfare can mitigate moral hazard effects in job search and wage formation (Andersen and Svarer 2014). It can work as a "threat/motivation" that makes unemployed individuals search harder and lower their reservation wages in order to get a job and thereby avoid program participation.

Workers that complete a program may also get better competencies that raise job finding rates and future earnings through a "program effect." On the other hand, job finding rates may decrease when participants are in the program because they have less time for job search or wish to complete the program—that is, a "lock-in effect."[8]

With all these hard-to-observe potential effects in play, it is difficult to estimate the benefits and costs of active labor market programs and how to make specific design decisions for these programs.

Lessons from the Ongoing Danish Policy Evaluation

For a country that uses as many resources on active labor market policy as Denmark, it is especially important to go beyond theory and build confidence in how different active labor market policies work and how to best allocate resources across different types of programs.

There has been a strong focus in the recent decades on evidence-based policymaking in Denmark's active labor market policy.[9] The goal is that decisions on how to design the policy and on the amount of resources to use rely as far as possible on cost-benefit analyses based on high-quality empirical evidence. This evidence is based partly on lessons from the international empirical literature (for example, as surveyed in Card, Kluve, and Weber 2018), on Danish register data-based

[8] On the macroeconomic level, the presence of activation may affect wages negatively or positively depending on its effect on the outside option of employed (wage effect), and it can induce more vacancy creation if aggregate search effort is increased, which also increases the benefit for firms of posting vacancies (vacancy effect). In addition, there may be spillover effects to other unemployed individuals if, for example, participating in an active measure increases job chances of treated unemployed individuals on behalf of job chances of untreated unemployed individuals competing for the same jobs (congestion effects) (Crépon et al. 2013; Ferracci, Jolivet, and van den Berg 2014; Gautier et al. 2018a).

[9] For more details see https://www.star.dk/en/evidence-based-policy-making/, where the evidence strategy is formulated. It consists of three strands: collect existing evidence about what works, innovate new evidence in relation to this, and communicate the results. The process of involving research in the actual policymaking has been a relatively long tradition in Danish labor market policy, and is presumably attributed to the early access to high quality micro data on individual labor market spells since the 1990s.

evaluations using modern identification strategies to identify causal effects, and on a long sequence of large-scale randomized control trial experiments organized by the Ministry of Labor. The systematic use of randomized control trials to evaluate the impact of the active labor market policies is a rather unique feature of the Danish labor market policy. The randomized control trials have the additional advantage that they provide a natural setting for evaluating the cost-effectiveness of the programs.

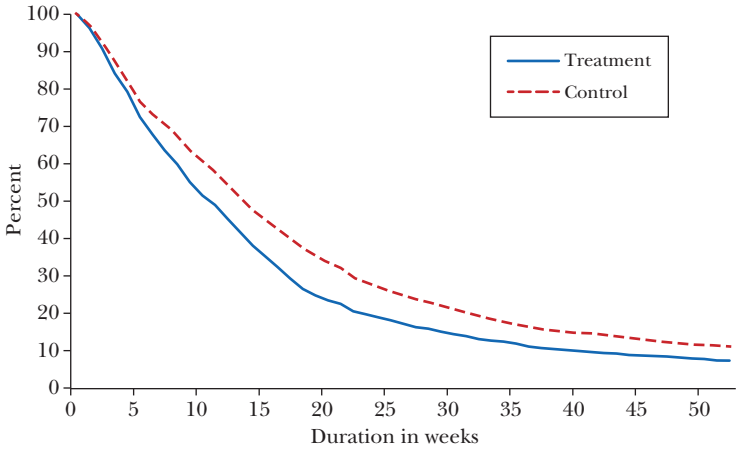
The Danish Ministry of Labor has organized eleven randomized experiments since 2005. The first experiment, called Quickly Back to Work, was conducted in two counties in Denmark during the winter of 2005–2006 and was targeted at newly unemployed recipients of unemployment insurance. All individuals in the two counties who became unemployed, and who were entitled to unemployment insurance benefits during this period, were allocated to either a treatment group or a control group. In practice, those born on the 1st to the 15th were given the treatment, and those born on the 16th to the 31st were not. The treatment consisted of intensified labor market measures, involving information, early mandatory participation in job search assistance programs, frequent meetings with case workers, and full-time program participation in an active labor market program for at least three months for those still unemployed after 18 weeks.

Figure 4 from an analysis by Gautier et al. (2018a) shows the unconditional effects on the employment status of individuals due to the experiment. The survival curves in the diagram show the duration of unemployment of the newly unemployed individuals in the treatment and control groups. After ten weeks, about half of the people in the treatment group have left unemployment, and half are still unemployed. The share still unemployed in the control group is around 60 percent. The 10 percentage-points lower unemployment rate in the treatment group corresponds to a reduction of 15 percent. The difference between the two groups widens up to around 20 weeks of elapsed duration. At this point, the number of people who are still unemployed is 30 percent lower in the treatment group compared to the counterfactual unemployment in the control group.

Several authors have evaluated this experiment in more detail. Graversen and van Ours (2008) apply duration models and find that the re-employment rate increases about 30 percent in the early phase of the unemployment period. Both Graversen and van Ours (2008) and Vikström, Rosholm, and Svarer (2013) investigate which elements of the activation program are most effective and find that the threat effect of activation and job search assistance are most effective. That is, unemployed respond to the requirement of participating in activation by leaving unemployment at an increasing rate as the time of activation is approaching. Rosholm (2008) finds that the estimated propensity to participate in meetings or being activated drives the difference in the job finding rates between treated and non-treated individuals. The Danish Economic Council (2007) has computed the impact on the government budget, including saved unemployment insurance benefits, of Quickly Back to Work to be a surplus of around 15,000 Danish kroner (approximately $2,500) per unemployed person in the experiment.

Figure 4

Exit from Unemployment for Treatment and Control Groups in the Danish Quickly Back to Work Experiment



Source: Gautier et al. (2018b).

Note: The figure partly replicates Figure 4 of Gautier et al. (2018a).The figure shows the fraction of individuals that are still unemployed at different elapsed durations of unemployment. The figure distinguishes between unemployed individuals that participated in a randomized controlled experiment (Quickly Back to Work) that offered more frequent participation in active labor market programs than the control group, which was subject to the traditional labor market policy. The graphs are based on weekly unemployment data for the two groups of unemployed.

The success in terms of positive effects on employment and public finances of Quickly Back to Work paved the way for further experiments that sought to disentangle the effects of the individual measures. A subsequent experiment implemented in 2008 separately studied the effects of three types of interventions: more frequent individual meetings with case workers; start of activation in job training/education after 13 weeks instead of after 26 weeks; and use of individual meetings versus cheaper group meetings with caseworkers. Maibom, Rosholm, and Svarer (2017) find that the treatment group accumulates more weeks in employment across all three interventions. In addition, all three interventions had a positive impact on public finances. The effect on public finances is best for individual meetings, then group meetings, and finally early activation.

The findings from the two experiments combined with supporting evidence from the economic literature have had a strong influence on Danish labor market policy, with early and frequent individual meeting activity of unemployed individuals with their caseworkers now being the norm.

In addition to the experimental evidence, microeconometric evidence on Danish population register data in Rosholm and Svarer (2008) shows a strong effect on the exit rate from unemployment even before the unemployed enter active labor

market policies. This evidence of a "threat effect" from active labor market policies aligns with evidence from other countries (Black et al. 2003; Hall et al. 2022) and suggests that the active labor market policies mitigate the adverse selection and moral hazard effects of high unemployment insurance benefits in line with the workfare theory of labor market policy.

Subsequent experiments have focused on unemployed individuals with a more marginal attachment to the labor market: for example, long-term social assistance recipients, people on sickness benefits (for example, Rehwald, Rosholm, and Rouland 2018), and young unemployed individuals with mental or cognitive challenges (for example, Rosholm, Mikkelsen, and Svarer 2019). The results from these experiments are less positive in terms of improving employment status and cost-effectiveness, and often they do not provide solid evidence for using active labor market measures for unemployed individuals with weak attachment to the labor market.

In summary, the benefits of workfare in active labor market policy is well founded in theory and evidence, although the effects do vary considerably across program characteristics and targeted groups. Indeed, a subset of the evaluated programs did not meet cost-effectiveness requirements, thereby pointing to the need for continuous evaluation and redesigning of active labor market policies.

The Challenges of Globalization, Automation, and Immigration

In recent decades, labor markets in many developed economies have been challenged by globalization, automation, and immigration. Although these developments are likely beneficial for aggregate income, they can also pose a disruptive threat for employment and especially for the income of low-skilled workers. Outsourcing of production to low-wage countries moves domestic low-skilled jobs away. Automation and the adoption of industrial robots reduce the demand for low-skilled labor. An inflow of foreign labor seeking employment opportunities may push down wages or employment prospects of native low-skilled individuals.

However, as is clear from Table 1, Denmark is doing quite well on measures of low unemployment, many workers in low-paying jobs, and a relatively equal distribution of incomes. One possibility is that the Danish economy is more isolated from these forces. The alternative is that the Danish labor market and flexicurity are performing well in accommodating the challenges.

Many facts suggest that the Danish economy is affected like other developed countries by globalization, automation, and immigration. Denmark is a small-open economy inside the European Union where agreements ensure free mobility of labor and capital. Denmark has a high degree of international collaboration and exchange of goods and service. For example, the foreign value added as a share of Danish exports is 33 percent, compared to an OECD average of 24 percent (OECD 2016).

Hummels et al. (2014) investigate the effects of offshoring by Danish manufacturing firms and find that offshoring leads to a reduction in employment,

primarily through a reduction in low-skill workers. In addition, offshoring increases wages of high-skilled workers, but decreases wages of low-skilled workers. Related, Utar (2018) investigates the effects of Chinese import penetration on workers in Danish firms and finds that workers exposed to competition face a higher risk of unemployment.

Acemoglu and Restrepo (2020) show that Denmark, in an international comparison, has a high adaption of industrial robots. For the United States, they find that the increased use of industrial robots reduces employment and wages in local labor markets. In a Danish context, Humlum (2019) finds that industrial robots have increased average real wages but lowered real wages of production workers employed in manufacturing. This can account for one-quarter of the fall in the employment share of production workers in Denmark since 1990.

In short, the Danish labor market seems strongly affected by globalization and automation, as are many other countries. However, the good Danish labor market performance indicates that the flexicurity model, with its massive spending in active labor market policy and education, appears to be accommodating the shocks and facilitating the necessary reallocation of labor.

Inflow of low-skilled immigrant labor may also pose a threat to native low-skilled workers, but this conclusion is not obvious. Foged and Peri (2016) find that an increase in the supply of refugee-country immigrants in Denmark pushed less-educated native workers, especially young and low-tenured workers, to pursue less manual-intensive occupations. As a result, and somewhat unexpectedly, immigration affected native unskilled wages and employment positively.

On the other hand, the Danish model does seems to have difficulties in integrating low-skilled immigrants into the labor market. The employment gap between natives and non-natives in Denmark is close to 30 percentage-points, which is higher than the OECD average and significantly higher than the US gap, which is below 20 percentage-points (OECD 2017c). One reason might be that the Danish minimum wages become an entry barrier for these individuals who do not have the same basic education background as the natives and do not share the language, thereby making integration policy more challenging.

Some Open Questions

The Danish flexicurity policy combines flexible hiring and firing rules of firms with high income security of low-skilled workers ensured by a long duration of unemployment insurance benefits and high income replacement rates. However, the key to the success of the Danish flexicurity policy since the early 1990s is its extensive use of active labor market policies, with participation in the programs being both a right and a duty of the unemployed. The workfare requirement to spend time in these programs tests the willingness to work of unemployed individuals and reduces the adverse selection and moral hazard problems of a high unemployment compensation. Denmark does not give up on getting an unemployed individual back into

work. The Danish active labor market programs are subject to ongoing microeconometric evaluation, building to a large extent on regular randomized control trial experiments. Reassuringly, the evidence shows that the policy has the intended effects, although effects vary across program characteristics and targeted groups.

For American readers, an obvious question is whether it is feasible and desirable for the United States to adopt its own version of Danish flexicurity. There are several difficult issues here.

First, the population of Denmark is similar to that of a single mid-sized American state like Colorado or Wisconsin. The Danish population is very homogenous and everyone receives, more or less, the same basic education in public schools. The problems with integration of immigrants into the Danish labor market suggest that it might be more difficult and expensive, or even infeasible, to implement nationwide flexicurity in a country with a more heterogeneous population such as the United States.

Second, a necessary condition for the successful combination of high income security and high employment is massive public spending on active labor market policy and, maybe, also spending on education in general. As noted earlier, Danish spending on active labor market policy alone corresponds to 2 percent of aggregate income, the highest level in the OECD, compared to 0.1 percent in the United States. The US GDP will probably exceed $22 trillion in 2022, and so spending 2 percent of that amount would be $440 billion per year. This amount does not include Denmark's high direct spending on unemployment compensation and income support. For a discussion of how it is possible to tax so much in countries like Denmark, a useful starting point is Kleven (2014) in this journal.

Third, practical implementation of Danish-style active labor market policies requires a substantial number of caseworkers who need to have a high level of government information about individuals, given that unemployed individuals are allocated to different active labor market policies based on discretionary assessments of caseworkers.

Finally, prevailing social attitudes about fairness of outcomes are important for how people view inequality and the need for policy action (Hvidberg, Kreiner, and Stantcheva 2021). A flexicurity policy is expensive for taxpayers and disproportionally helps people with weak labor market attachments. Scandinavians are more likely to perceive these individuals as being unlucky, rather than lazy, and as having small chances of upward mobility compared to Americans (Alesina, Glaeser, and Sacerdote 2001; Alesina, Stantcheva, and Teso 2018). Danes also seem to have a higher trust in government and stronger civic virtues (row 3 in Table 1). For example, Algan and Cahuc (2009) measure civic-mindedness based on survey responses to this question: "Do you think it can always be justified, never be justified, or something in between to claim government/state benefits to which you have no rights." By this measure, Denmark leads the countries in this sample on civic-mindedness, while the US responses are in the middle of the pack. A high degree of civic-mindedness in this sense can both make unemployed workers more responsive to active labor market programs and help to create broad-based political support for a flexicurity policy.

■ *We are grateful to Torben M. Andersen, Richard Blundell, Gordon Dahl, Daniel le Maire, and the editors, Erik Hurst, Nina Pavcnik, Timothy Taylor, and Heidi Williams, for helpful discussions and suggestions. We thank Simon Kyllebæk Andersen for outstanding research assistance and Martin Ulrik Jensen, Mads Kieler and Nina Marquardt from the Danish Ministry of Finance, Louise Broman and Lasse Bank from the Danish Ministry of Employment, and Andrea Salvatori from the OECD for assistance with data. Kreiner gratefully acknowledges funding from the Danish National Research Foundation (grant DNRF134).*

References

Acemoglu, Daron, and Pascual Restrepo. 2020. "Robots and Jobs: Evidence from US Labor Markets." *Journal of Political Economy* 128 (6): 2188–244.

Alesina, Alberto, Edward Glaeser, and Bruce Sacerdote. 2001. "Why Doesn't the United States Have a European-Style Welfare State?" *Brookings Papers on Economic Activity* 2: 187–278.

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso. 2018. "Intergenerational Mobility and Preferences for Redistribution." *American Economic Review* 108 (2): 521–54.

Algan, Yann, and Pierre Cahuc. 2009. "Civic Virtue and Labor Market Institutions." *American Economic Journal: Macroeconomics* 1 (1): 111–45.

Andersen, Torben M. 2019. "The Danish Labour Market, 2000–2018." *IZA World of Labour*, December. https://wol.iza.org/uploads/articles/514/pdfs/the-danish-labor-market.pdf.

Andersen, Asger Lau, Amalie Sofie Jensen, Niels Johannesen, Claus Thustrup Kreiner, Søren Leth-Petersen, and Adam Sheridan. 2021. "How Do Households Respond to Job Loss? Lessons from Multiple High-Frequency Data Sets." CEPR Discussion Paper 16131.

Andersen, Torben M., and Michael Svarer. 2007. "Flexicurity: Labour Market Performance in Denmark." *CESifo Economic Studies* 53 (3): 389–429.

Andersen, Torben M., and Michael Svarer. 2014. "The Role of Workfare in Striking a Balance Between Incentives and Insurance in the Labour Market." *Economica* 81 (321): 86–116.

Besley, Timothy, and Stephen Coate. 1992. "Workfare versus Welfare: Incentive Arguments for Work Requirements in Poverty-Alleviation Programs." *American Economic Review* 82 (1): 249–61.

Besley, Timothy, and Stephen Coate. 1995. "The Design of Income Maintenance Programmes." *Review of Economic Studies* 62 (2): 187–221.

Bick, Alexander, Bettina Brüggemann, and Nicola Fuchs-Schündeln. 2019. "Hours Worked in Europe and the United States: New Data, New Answers." *Scandinavian Journal of Economics* 121 (4): 1381–416.

Black, Dan A., Jeffrey A. Smith, Marc C. Berger, and Brett J. Noel. 2003. "Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence from Random Assignment in the UI System." *American Economic Review* 93 (4): 1313–27.

Boeri, Tito, Agar Brugiavini, and Lars Calmfors. 2001. "The Role of Unions in the Twenty-First Century." Oxford: Oxford University Press.

Card, David, Jochen Kluve, and Andrea Weber. 2018. "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations." *Journal of the European Economic Association* 16 (3): 894–931.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 128 (2): 531–80.

Dahl, Christian M., Daniel le Maire, and Jakob R. Munch. 2013. "Wage Dispersion and Decentralization of Wage Bargaining." *Journal of Labor Economics* 31 (3): 501–33.

Danish Economic Council. 1995. *Report on the Danish Economy, Spring 1995*. Danish Economic Council: Copenhagen.

Danish Economic Council. 2007. *Report on the Danish Economy, Spring 2007*. Danish Economic Council: Copenhagen.

Danish Economic Council. 2014. *Report on the Danish Economy, Autumn 2014*. Danish Economic Council: Copenhagen.

Ferracci, Marc, Grégory Jolivet, and Gerard J. van den Berg. 2014. "Evidence of Treatment Spillovers within Markets." *Review of Economics and Statistics* 96 (5): 812–23.

Foged, Mette, and Giovanni Peri. 2016. "Immigrants' Effect on Native Workers: New Analysis on Longitudinal Data." *American Economic Journal: Applied Economics* 8 (2): 1–34.

Gautier, Pieter, Bas van der Klaauw, Paul Muller, Michael Rosholm, and Michael Svarer. 2018a. "Estimating Equilibrium Effects of Job Search Assistance." *Journal of Labor Economics* 36 (4): 1073–125.

Gautier, Pieter, Bas van der Klaauw, Paul Muller, Michael Rosholm, and Michael Svarer. 2018b. "Estimating Equilibrium Effects of Job Search Assistance." Journal of Labor Economics 36 (4): 1073–125. Unpublished data (accessed June 2021).

Graversen, Brian Krogh, and Jan C. van Ours. 2008. "How to Help Unemployed Find Jobs Quickly: Experimental Evidence from a Mandatory Activation Program." *Journal of Public Economics* 92 (10-11): 2020–35.

Hall, Caroline, Kaisa Kotakorpi, Linus Liljeberg, and Jukka Pirttilä. 2022. "Screening through Activation? Differential Effects of a Youth Activation Program." *Journal of Human Resources 57 (3):* 1033–1077.

Hansen, Claus Thustrup, and Torben Tranæs. 1999. "Effort Commitment in Active Labour Market Policy." In *Macroeconomic Perspectives on the Danish Economy*, edited by Torben M. Andersen, Svend E. Hougaard Jensen, and Ole Risager, 335–64. Macmillan Press: London.

Hermansen, Mikkel. 2015. "Evaluering af dagpengereformen: Beskæftigelseseffekter ved halveringen af dagpengeperioden" (Evaluation of the unemployment insurance reform: Employment effects from reducing the benefit duration period). *Danish Economic Journal* 2015 (1): 1–35.

Høgedahl, Laust. 2020. "The Danish Labour Market Model: Is the Bumblebee Still Flying?" In *Oxford Handbook of Danish Politics*, edited by Peter Munk Christiansen, Jørgan Elklit, and Peter Nedergaard, 559–76. Oxford: Oxford University Press.

Humlum, Anders. 2019. "Robot Adoption and Labor Market Dynamics." Unpublished.

Humlum, Anders, and Jakob Roland Munch. 2019. "Globalization, Flexicurity and Adult Vocational Training in Denmark." In *Making Globalization More Inclusive: Lessons from Experience with Adjustment Policies*, edited by Marc Bacchetta, Emmanuel Milet, and José-Antonio Monteiro, 51–69. Geneva, Switzerland: World Trade Organization.

Hummels, David, Rasmus Jørgensen, Jakob R. Munch, and Chong Xiang. 2014. "The Wage Effects of Offshoring: Evidence from Danish Matched Worker-Firm Data." *American Economic Review* 104 (6): 1597–629.

Hvidberg, Kristoffer B., Claus Thustrup Kreiner, and Stefanie Stantcheva. 2021. "Social Positions and Fairness Views on Inequality." NBER Working Paper 28099.

International Labour Organization (ILO). 2019. *Quick Guide on Interpreting the Unemployment Rate.* Geneva, Switzerland: International Labour Organization.

Immervoll, Herwig, Henrik Jacobsen Kleven, Claus Thustrup Kreiner, and Emmanuel Saez. 2007. "Welfare Reform in European Countries: A Microsimulation Analysis." *Economic Journal* 117 (516): 1–44.

Jonassen, Anders B. 2013. "Regression Discontinuity Analyses of the Disincentive Effects of Increasing Social Assistance." PhD diss. Aarhus University.

Kleven, Henrik Jacobsen. 2014. "How Can Scandinavians Tax So Much?" *Journal of Economic Perspectives* 28 (4): 77–98.

Kleven, Henrik Jacobsen, and Claus Thustrup Kreiner. 2005. "Labor Supply Behavior and the Design of Tax and Transfer Policy." *Danish Economic Journal* 143 (2005): 321–58.

Kleven, Henrik Jacobsen, Martin B. Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez. 2011. "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark." *Econometrica* 79 (3): 651–92.

Kleven, Henrik Jacobsen, Claus Thustrup Kreiner, and Emmanuel Saez. 2016. "Why Can Modern Governments Tax So Much? An Agency Model of Firms as Fiscal Intermediaries." *Economica* 83: 219–46.

Kreiner, Claus Thustrup, Daniel Reck, and Peer Ebbesen Skov. 2020. "Do Lower Minimum Wages for Young Workers Raise Their Employment? Evidence from a Danish Discontinuity." *Review of*

Economics and Statistics 102 (2): 339–54.

Kreiner, Claus Thustrup, and Michael Svarer. 2022. "Replication data for: Danish Flexicurity: Rights and Duties." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E174561V1.

Kreiner, Claus Thustrup, and Torben Tranæs. 2005. "Optimal Workfare with Voluntary and Involuntary Unemployment." *Scandinavian Journal of Economics* 107 (3): 459–74.

Maibom, Jonas, Michael Rosholm, and Michael Svarer. 2017. "Experimental Evidence on the Effects of Early Meetings and Activation." *Scandinavian Journal of Economics 119 (3): 541–70.*

McDonald, Ian M., and Robert M. Solow. 1981. "Wage Bargaining and Employment." *American Economic Review* 71 (5): 896–908.

Miller, Terry, Anthony B. Kim, and James M. Roberts. 2019. *2019 Index of Economic Freedom.* Washington, DC: The Heritage Foundation. https://www.heritage.org/index/excel/2019/index2019_data.xls (accessed September 2, 2021).

Ministry of Employment. 1993-2019. "Data for Aggregate Activation Rates." Unpublished data (accessed June 2021).

OECD. 2016. *OECD Factbook 2015–2016: Economic, Environmental and Social Statistics.* Paris: OECD Publishing.

OECD. 2017a. "Social Expenditure – Aggregated data." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=SOCX_AGG (accessed April 21, 2021).

OECD. 2017b. "Educational Finance Indicators: C3.1: Relative share of public, private and international expenditure on educational institutions, by final source of funds." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=EAG_FIN_RATIO (accessed May 5, 2021).

OECD. 2017c. *Employment Outlook 2017.* Paris: OECD Publishing.

OECD. 2018a. "Income Distribution Database." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=IDD (accessed April 14, 2021).

OECD. 2018b. "OECD Employment Outlook 2018." OECD Publishing, Paris. https://doi.org/10.1787/888933778212%20 (accessed May 19, 2021).

OECD. 2018c. "Public Expenditure and Participant Stocks on LMP: Public Expenditure of LMP by Main Categories (% GDP)." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=LMPEXP (accessed April 8, 2021).

OECD. 2018d. "Trade Union Dataset." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=TUD (accessed April 8, 2021).

OECD. 2019a. "Better Life Index." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=BLI (accessed April 9, 2021).

OECD. 2019b. "Government at a Glance 2019." OECD Publishing, Paris. https://doi.org/10.1787/888934033137 (accessed July 18, 2022).

OECD. 2019c. "Disposable Income and Net Lending – Net Borrowing." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=SNA_TABLE2 (accessed April 14, 2021).

OECD. 2019d. "Gross Domestic Product (GDP)." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=SNA_TABLE1 (accessed April 9, 2021).

OECD. 2019e. "Decile Ratios of Gross Earnings: Incidence of Low Pay." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=DEC_I (accessed April 15, 2021).

OECD. 2019f. "Short-Term Labour Market Statistics." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=STLABOUR (accessed April 9, 2021).

OECD. 2019g. "Incidence of Unemployment by Duration." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=DUR_I (accessed April 9, 2021).

OECD. 2019h. "Short-Term Labour Market Statistics: Unemployment Rates by Age and Gender." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=STLABOUR (accessed April 9, 2021).

OECD. 2019i. "Revenue Statistics – OECD Countries: Comparative Tables." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=REV (accessed April 14, 2021).

OECD. 2019j. "Social Expenditure – Aggregated data." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=SOCX_AGG (accessed April 14, 2021).

OECD. 2019k. "Net Replacement Rate in Unemployment." OECD.Stat. https://stats.oecd.org/Index.aspx?DataSetCode=NRR (accessed on April 22, 2021).

OECD. 2020a. "OECD Employment Protection Legislation Database, 2020 Edition." OECD. https://www.oecd.org/els/emp/OECDEmploymentProtectionLegislationDatabase.xlsx accessed September 1, 2022).

OECD. 2020b. "The Design of Unemployment Benefits Schedules over the Unemployment Spell: The Case of Belgium." OECD Social, Employment and Migration Working Paper 2020-2. Unpublished data (accessed April 30, 2021).

OECD. 2020c. *The OECD Tax-Benefit Model for Denmark*. Paris: OECD Publishing.

OECD. 2020d. *The OECD Tax-Benefit Model for The United States*. Paris: OECD Publishing.

OECD. 2020e. *TaxBEN: The OECD Tax-Benefit Simulation Model*. Paris: OECD Publishing.

Pedersen, Peder J., and Nina Smith. 1995. "Search Activity among Employed and Unemployed Members of the Workforce." In *Work Incentives in the Danish Welfare State*, edited by Gunnar Viby Mogensen. Aarhus: Aarhus Universitet.

Rehwald, Kai, Michael Rosholm, and Bénédict Rouland. 2018. "Labour Market Effects of Activating Sicklisted Workers." *Labour Economics* 53 (C): 15–32.

Rosholm, Michael. 2008. "Experimental Evidence on the Nature of the Danish Employment Miracle." IZA Discussion Paper 3620.

Rosholm, Michael, and Michael Svarer. 2008. "Threat Effect of Active Labour Market Programs." *Scandinavian Journal of Economics* 110 (2): 385–401.

Rosholm Michael, Mai Mikkelsen, and Michael Svarer. 2019. "Bridging the Gap from Welfare to Education: Propensity Score Matching Evaluation of a Bridging Intervention." *PLoS ONE* 14 (5).

Statistics Denmark. 1980–2020. "ADAM Data Bank of Statistics Denmark." Unpublished data.

Svarer, Michael. 2011. "The Effect of Sanctions on Exit from Unemployment: Evidence from Denmark." *Economica* 78 (312): 751–78.

Utar, Hale 2018. "Workers Beneath the Floodgates: Low-Wage Import Competition and Worker's Adjustment." *Review of Economics and Statistics* 100 (4): 631–47.

Vikström, Johan, Michael Rosholm, and Michael Svarer. 2013. "The Relative Efficiency of Active Labour Market Policies: Evidence from a Social Experiment and Non-Parametric Methods." *Labour Economics* 24: 58–67.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 103–124

Debt Revenue and the Sustainability of Public Debt

Ricardo Reis

At the end of 2020, gross US government debt was 134 percent of GDP, the highest in US history, well above its previous record (121 percent just after World War II in 1946). The records for the size of public debt have likewise been broken for the groups of advanced economies and of emerging market economies (IMF 2021a). This was not solely the result of the pandemic, because debt had been growing since the 1980s and at a rising pace since the great financial crisis of 2008–2009 (Yared 2019). Is this level of debt sustainable, both for the US economy and for others around the world?

Governments have had centuries of experience actively using the public debt to prevent sharp changes in taxes or spending. Sometimes they just passively roll the debt over for many years, hoping for the best or falling for the seduction of reckless schemes. Economic theorists have analyzed how much and for how long debt can be sustained using impressive-sounding concepts like "bubbles," "Ponzi schemes," and "transversality conditions." Together, theory and experience have shown that ever-delaying the collection of taxes to pay for past debts is sometimes possible, but always eventually limited. Recently, a growing literature has found a third method by which to sustain public debt: to collect some new revenue every time that new public debt is issued. I call this the *debt revenue*. This essay describes where it comes from and its implications for whether the current level of public debt is sustainable.

■ *Ricardo Reis is Professor of Economics, London School of Economics, London, United Kingdom. His email address is r.a.reis@lse.ac.uk.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.103.

What is debt revenue? When the government tries to sell a public bond, it must compete with many other prospective borrowers, including foreign governments, firms, and even households, as a bank that lends more to the government may cut back on its personal credit. There is a market interest rate at which the borrowing by all equals the total amount lenders are willing to give. For some reason, the creditors give the government a discount, charging less on the public debt than that market rate. This discount times the amount of debt is the debt revenue. It saves the government the need to collect future taxes to repay a debt that grows at a lower rate than market returns.

Many governments in the past two decades received such a large discount that the real interest rate they paid was negative. In these cases, the revenue is visible: creditors give more today than what the government will pay them back in the future, so the government can set aside the repayment and spend the difference right away. But even if the real interest rate is positive, there is a debt revenue as long as there is a discount. The revenue may be realized, if the government borrows at the reduced rates and gives public loans at close-to-market rates, keeping the profits. Or, it may be implicit, by considering hypothetical counterfactuals: the government could borrow at its discounted rate, transfer that amount to households that were previously borrowing at market rates, and later tax those same households back by the original amount lent times the market rate. The household's resources have not changed at all, but the government is left with the debt revenue after it pays the original government debt. Another way to see the debt revenue is through the lens of the sustainability of public finances: for a given plan for spending and taxes, the public debt will grow at a slower rate as a result of the discount; without it, debt would explode faster and require that austerity arrives sooner.

Why has this debt revenue been negligible, and so typically ignored, in analyses of debt sustainability? What is special about government debt that gives rise to the discount in the returns that it pays its creditors in the first place? How large is the debt revenue, and how does it compare with the seignorage that central banks earn, a more familiar revenue from issuing a public liability? Does debt revenue come with different trade-offs facing policymakers when deciding how much to spend and tax? This article reviews the answers that a rapidly growing literature has given to these questions.[1]

Classic Analyses of Debt Sustainability[2]

The definition of debt sustainability has one equation at its foundation: the government budget constraint.

[1] Willems and Zettelmeyer (2022) provide a complementary review.

[2] Recent excellent examples of the classical analysis are in Gale (2019), Abbas, Pienkowski, and Rogoff (2020), and Eichengreen et al. (2022).

The Classic Version of the Government Budget Constraint

It is an accounting identity that, for a given year:

Increase in public debt = return to debtholders - primary balance.

The primary balance is the difference between tax revenues and government spending (on purchases and transfers). When it is positive, there is a surplus, and when it is negative, a deficit. The return to debtholders is the sum of: the promised interest rates on the debt, the repayment of the amount borrowed for debt that is coming due, and the change in the value of debts that will only come due in the future. The sum of balance and returns gives the left-hand side: the increase in the market value of the debt. As an identity, this equation always holds.

Starting from the market value of debt today, the equation tells us what debt will be next year. The same applies to the next year, the year after, and so on, linking today's market value of debt to what its value might be in a far-away future. However, the future balances are in the units of goods in the future, and the value of goods in the future is not the same as their value today. To add up these increments, one has to multiply the market values at future dates by their price in today's units. This can be expressed as the future increments to the debt being discounted at a rate d, as payments in the future are typically worth less than resources right away.

For decades, economists chose d to be the returns on government debt r. This seems like a natural discount rate for future deficits because it is the rate at which the government borrows to roll over pre-existing debt. This choice gives rise to the equation:

$$\frac{Debt}{GDP} = EPV_{r-g}\left(\frac{PrimaryBalance}{GDP}\right).$$

The notation $EPV_{r-g}(.)$ stands for the expected present value, using the returns on public debt r that are paid by government as the discount rate. Scaling by GDP is important because taxes and government spending can only be as large as the size of the economy. A positive debt has to be paid with positive surpluses at some point in the future, but these may be either a negligible or a significant share of the economy's income that year.

This equation is identical to the accounting identity as long as one important condition is met: that, on average and over the distant future, r is larger than the growth rate of the economy g. Otherwise, because the primary balances are growing with the economy, the future increments are growing faster than they are being discounted. The right-hand side would not be properly defined. If $r > g$ though, this is still just an accounting identity, even if now written in an intertemporal form (mathematically, it is the integral form of the differential equation).

Traditional Debt Sustainability

The equation relates the value of the public debt, on the left-hand side, to the expected present value of the balances that the government will use to pay it down, on the right-hand side. It is just like the relation between a stock price and expected present value of dividends. If the balances are too low, then the market value of the debt will be low, and the investors that held the debt from the past will be making losses on these holdings. If the expected present value of the balances is so low that it is below the payments on the debt that are due today, then the government has no choice but to default, paying back less than what was promised. More generally, the public debt is unsustainable if there needs to be either a default (so the left-hand side falls) or a reversal in public finances that generates a large increase in future primary balances (so the right-hand side rises) in order to bring the two sides in line.

Assessing the size of the right-hand side of the equation and comparing it with the size of the debt that is due or outstanding gives an assessment of debt sustainability. Different lenders may have different perspectives on what the future will be and may change their minds suddenly. When they do, the value of the debt can change suddenly, so the government wants to anticipate these changes with its own estimates. Fiscal councils (like the Congressional Budget Office in the United States) can perform a useful role in providing credible estimate of the right-hand side to inform and anchor the market expectations. If that estimate is well below the current market value of the debt, there are reasons to be concerned, as a sharp drop in government bond prices may be on the horizon.

Measuring public debt on the left-hand side of the equation may seem easy, but in practice a comprehensive measure can be tricky. First, it is important to add and subtract the debt issued and held by different branches of the state, including regional governments and social security trust funds. Second, and more difficult, one should subtract from the state's liabilities the assets that it will be willing to sell if needed to honor the public debt. Third, and very hard, the measurement has to depend on what will happen in the future since, in times of crisis, public debt can jump when certain contingencies are triggered as governments take on commitments (like insuring mortgages or business loans).

Measuring Future Primary Balances

A popular way to measure the right-hand side is to build forecasts of future primary balances. Experience has shown that the uncertainty around these forecasts is very large. For example, small differences in plausible assumptions about retirement age, the cost of health care, and what future governments will choose to tax during the next couple of decades can produce forecasts that differ by several multiples. With an $r - g$ that is around 0.01 or 0.02, budget forecasts that are as far away as 50 years from now will still loom quantitatively large in the present value calculations. A more fundamental challenge is that, since all countries have positive debt, sustainability suggests that these forecasts must include positive balances sometime in the future. However, for many countries, and the United States in particular,

the forecasts are for deficits for the foreseeable future. When the IMF conducts an analysis of debt sustainability as part of its surveillance of member countries, eventually, even if in a distant future, it always assumes that primary balances become positive.

These difficulties have led to a second approach to measure debt sustainability. It asks a slightly different question: whether, after a sudden shock to the public finances that raises the public debt, this new debt will be paid for by future balances. The answer turns out to be simple. As long as an increase in public debt leads eventually to an increase in the primary balance, even if only in the distant future, the debt is sustainable. Using past data to estimate how fiscal policy, through rules, common practices, or discretionary choices, changed primary balances in response to higher debt provides an indication of whether it will do so in the future. These responses of primary balances to past public debt are called "fiscal reaction functions." Estimates using data for advanced economies in the twentieth century have typically found a positive relation, leading to the conclusion that debt was always sustainable (Bohn 1998). At the same time, econometric identification of these rules is challenging, and the exercise makes the strong assumption that past patterns of fiscal policy reflect its future behavior.

A third approach to assess debt sustainability is to calculate the feasible maximum value of the right-hand side of the government budget constraint. Instead of trying to figure out what the government will do, it calculates what is the most that the goverment can feasibly do. If that is less than the outstanding debt, then the debt is unsustainable. To calculate the maximum requires models (D'Erasmo, Mendoza, and Zhang 2016). Most of them are versions of Laffer curves: relations between tax revenues and tax rates. Higher rates raise revenues at first, but eventually higher rates may discourage the desire to work, to invest, or to comply with the tax authorities, so that revenues actually fall. This peak of the curve gives the maximum revenue that the government can collect to pay for its public debt. An important limitation of these analyses is that there is no corresponding analysis of the feasibility of cutting government spending, so that at least half of the components of the balance is left out.

Classic Trade-Offs

Perhaps the most famous trade-off in debt sustainability analysis is the one surrounding austerity. Cuts in spending and rises in taxes raise the primary budget balance, but they may also lower the growth rate of the economy, therefore increasing the rate at which these balances are discounted. Austerity that causes a recession may then actually lower the right-hand side and make public finances less sustainable (Alesina, Favero, and Giavazzi 2019).

A related trade-off arises from structural reforms. On the one hand, they are meant to raise the growth rate of the economy, and a higher g would raise the present value of primary surpluses. On the other hand, such reforms may require deficits at first to make the needed investments. Whether the right-hand side rises or falls depends on the relative weights of the present versus the future and on the

success of the uncertain outcome of the reforms (Müller, Storesletten, and Zilibotti 2019).

Another prominent trade-off arises in discussions of whether to default on public debt. A default would lower the left-hand side automatically. If, however, the holders of government debt suspect that a default is likely, they will require a higher return r. This raises the discounting of future surpluses, and so lowers the right-hand side. In this framework, sovereign debt crises can arise suddenly and there may be multiple equilibria (Calvo 1988).

A final trade-off involves inflation, which affects debt sustainability through three channels. First, unexpected inflation lowers the value of public debt on the left-hand side. Second, fully expected inflation has no effect on either side, as it raises r and g by the same amount. Third, higher risk of inflation raises r because investors require higher expected returns to hold a bond that may be debased by inflation in the future, so it lowers the right-hand side. In practice, bouts of inflation have unexpected, expected, and risk elements. Complicating matters further, historically, inflation often comes with financial repression that keeps r low and increases primary balances. An extreme example of financial repression is for the debt to be paid back with reserves at the central bank that pay zero interest, yet must be held for a long period of time (Hilscher, Raviv, and Reis 2022).

These trade-offs are interrelated, and more could be added. Together with the measurement of sustainability, they have led to an enormous literature in economics that has sought to provide guidance to policymakers.

The Debt Revenue

A remarkable fact of the first two decades of the twenty-first century is the steady decline in the real return on public debt (r minus inflation). In the United States, for instance, on average between 2001 and 2020, that real return was 2.5 to 3.5 percent lower (depending on the measure used) than in the preceding 20 years. Even before, throughout the nineteenth and twentieth centuries, the United States had the enviable position of paying a return on its government debt that was on average lower than the growth rate of income. Over the last 20 years, this gap has become larger, but also more widely shared across countries (Blanchard 2019; Mehrotra and Sergeyev 2021). As a consequence, the equation on which the traditional analysis of debt sustainability was sustained is no longer valid. Setting the appropriate discount rate d equal to the return r is no longer tenable because the expected present value is not well defined, diverging to infinity.

However, there is a sensible alternative: the return on private investments, call it m. The private sector as a whole can hold as assets either the government debt or the economy's capital. The return on private investment (the marginal product of capital) is the opportunity cost of holding the debt. At the margin, for investors to hold government debt, they must calculate the expected present value of payoffs from government bonds using the return on holding the capital stock. Moreover,

even as r declined in the last 20 years, m did not, staying approximately constant and well above the growth rate of the economy g.

Using m as the discount rate changes the government budget constraint described earlier. The public debt must still be backed by the present value of primary balances; the only change to the first term on the right-hand side of the equation is that the present value is now discounted by $m - g$. The same measurement difficulties and associated policy trade-offs apply to this term as they did in the classic analysis. But now there is a new positive term relative to traditional analysis, the debt revenue term (Reis 2021; Cochrane 2021). This term takes the debt/GDP ratio every year moving forward, multiplies it by $m - r$, and then calculates the expected present value.

$$\frac{Debt}{GDP} = EPV_{m-g}\left(\frac{Primary\ Balance}{GDP}\right) + EPV_{m-g}\left(\frac{(m-r)Debt}{GDP}\right).$$

This new equation is well defined even as r is less than g, and classic analyses of debt sustainability apply all the same to the first term. Moreover, when the return on government bonds and the return on private capital are the same ($m = r$), then the two equations are the same: the debt revenue term is equal to zero, and there is a single return to discount the future. There is even an a priori argument for why it *should* be so. If the return on private capital was higher than the return on government bonds, then private investors should invest more in the capital stock and less in government bonds. In doing so, the forces of demand and supply should make m fall and r rise until they are the same. However, this is not so in the data. As a result, the government earns a debt revenue.

Why Is There a Debt Revenue?

Since, for some reason, people are willing to hold public debt in spite of it giving a lower return than the private market alternative, their opportunity cost of doing so is a form of revenue for the government. Supplying the public debt is providing some service to these investors. The government is rewarded for it by being allowed to borrow at a lower interest rate than it otherwise would. The gap $m - r$ measures the discount that the government receives on the terms of its borrowing in exchange for these services.

Multiplying the premium by the total debt supplied gives the debt revenue flow. In any given year, this may be positive or negative. After all, private capital sometimes gives unusually higher returns because the economy boomed, but other times markets crash and the return on private investment can fall below that on government debt. Likewise, public debt can sometimes give unusually low returns because the government defaulted or because unexpectedly high inflation subtracted from the low nominal interest rates at which the debt had been sold. It is important to take the expected present value of the debt revenue flow to get to the value of the debt revenue and adjust for the riskiness of the debt revenue flows.

Perhaps the last two decades of ultra-low real returns on government bonds were just a very unusual random draw. Recently, the runaway inflation in 2021 and 2022

in many advanced economies has led to record low returns on government bonds, as the nominal payments they make to the bondholders are worth less in units of goods. Maybe as lenders start expecting inflation, they will require higher returns to lend to the government, so that r is about to rise to become again close to or equal to m. In other words, the flow of debt revenues of the last 20 years may have been a fluke, so their expected present value looking forward may be close to zero.

To figure out if it is so requires understanding what creates the debt revenue in the first place. It can only sustain public debt systematically as long as, on average, the return on private assets is expected to be higher than the return on government debt. Economists sometimes call this expected gap a *premium*. Something must impede the market forces that drive the premium to zero. Or, equivalently, there must be something special about government debt, or some unique service that it provides, to those who are willing to hold it. The literature has provided several arguments for what this might be (Krishnamurthy and Vissing-Jorgensen 2012).

Where Does the Premium between Returns Come From?

First, public debt is useful as a store of value that fills some holes left by the limitations of private credit markets. A primary function of credit is to allow resources to flow from the many who have them to the few who right now have an entrepreneurial project or an investment idea. However, the inability to sort out good from bad projects, or for the borrower to commit to repay, may put limits on this flow, leaving too many savers unable to put their resources to good use. At the same time, prospective lenders may have their own investment opportunities in the future, so they would like to save for the future. Public debt becomes useful because it provides an alternative store of value to the private credit markets that absorbs this excess supply of savings. The $m - r$ premium emerges because even though savers would like to put their capital to use in firms to earn m, the limits to private credit hinder this action, thus creating a residual demand for public debt even if at a lower r. Closely related, when there no better ways to store value, there may be a bubble raising the price of government bonds because some investors buy them expecting the price to keep on rising and returns to be high.[3]

Second, public debt is a safe haven. Holding a government bond bears the risk of unexpected movements in inflation, but the return on private investment is affected not just by inflation but also by almost any other shock to sales, investment, labor costs, or productivity. Moreover, loans to private firms are more likely to be defaulted on than loans to the government. The investors who want a safe asset are willing to pay for it by requiring a lower return in their loans to the government. That individual investment projects come with risks that are specific to the project is also relevant. Because much of this risk cannot be diversified away, firms and households would like to hold some of their savings aside in a safe asset. Finally, when uncertainty rises, investors fly to the safety of government bond from all other

[3]For some models of this, see Reis (2021), Miao and Su (2021), Bayer, Born, and Luetticke (2020), Bonam (2021), and Gersbach, Rochet, and von Thadden (2022).

assets and markets. This makes the returns on government debt rise during crises, which in turn makes investors tolerant of its low returns most of the time.[4]

Third, the premium may reflect regulations and financial repression. Many financial institutions are required by regulations to hold government bonds as a share of their assets or as collateral in some transactions. Governments routinely restrict households and institutions from some private-sector investment choices and put limits on private credit. This reduces the demand for the capital stock and correspondingly raises the demand for government bonds as an alternative, thus contributing the premium between their returns. In this case, the premium is akin to a repression tax, and the debt revenue is a tax revenue charged on private agents who are forced to lend to the government at inferior returns. Even when the government is not involved, lenders often repress their willingness to lend by the collateral that they are willing to accept from borrowers in the event that they default on their obligation. An asset is good for collateral if it is itself unlikely to default, if it is liquid so the lender can sell it easily, and if it is insensitive to new information so the lender does not need to spend resources keeping track of its value. Public debt is a natural candidate, and government bonds are used as collateral throughout the financial system. The premium then reflects this demand for collateral from households and firms.[5]

Fourth, government bonds are traded in liquid markets. This makes them easy to sell for cash and goods when their holders want to quickly increase their spending in goods. Most private investments, instead, take time to unwind or are difficult to sell because buyers are suspicious that the motive behind the sale is there being something wrong with the project. The gap between returns is referred to as a liquidity premium.[6]

There are many models in the economic literature to justify stores of value, safety, repression, or liquidity. A catch-all term that is sometimes used for all of them is to say that government bonds provide a convenience premium. Public debt, somehow, provides a convenience service to its holders. This special service is reflected in the low return on government bonds, and its associated revenue is captured by its issuer, the government.

Different Types of Public Debt and Seignorage

Government debt takes different forms. Some of it is in bonds that pay their holders a set amount of currency; some has its payments automatically rise with inflation. Some debts make small payments every year for a fixed set of years, others repay the creditor once just three months after they were paid for. These features determine how safe or liquid they are, so the premium on their returns varies, as

[4]Models of this channel are in Bassetto and Cui (2018), Bassetto and Cui (2021), Reis (2021), Elenev et al. (2021), Brunnermeier, Merkel, and Sannikov (2022), and Jiang et al. (2022).

[5]For models of this channel, see Angeletos, Collard, and Dellas (2016), Miao and Su (2021), and Gorton and Ordoñez (2022).

[6]Models of this liquidity are, for instance, in Berentsen and Waller (2018), Sims (2020), and Schmid, Liu, and Yaron (2021).

does the debt revenue the government earns from issuing them. All countries have government debt offices that vary the composition of the public bonds that they issue in part to try to maximize the debt revenue that is earned.

Central banks earn a particular form of revenue that is familiar to economists—seignorage—and which is closely related to debt revenue. When the central bank issues currency (for example, in the form of banknotes), the bank can buy goods with it. Seignorage is the change in that currency divided by the price of the goods. The central bank could instead use the newly printed banknotes to buy government bonds, and in fact this is what is usually done. The returns paid on those bonds to the central bank are then sent back to the government, so this becomes a form of debt revenue that can be used for government purchases or transfers every year from then onwards. The once-off seignorage is equal to the expected present value of the debt revenue from currency.

Monetary theory has for decades developed justifications for why people hold currency when it gives an inferior return to government bonds. It uses arguments on storage of value, risk, liquidity, and repression that mirror the ones made for public bonds above. However, in almost all advanced economies, the stock of currency is typically between 5 percent and 15 percent of GDP, and nominal returns have been close to zero, so seignorage has been trivial. In the history of the United States, the seignorage generated by the Federal Reserve has only very rarely been above 1 percent of real GDP in any one year, playing no meaningful role in directly sustaining the public debt. As the next section will show, the debt revenue from US Treasuries is an order of magnitude larger than that from currency, as public debt exceeds 100 percent of GDP and the gap between returns on private investment and on government bonds is several percentage points. Seignorage is one particular component of debt revenue, but one that is not particularly large.

Why Rethink Debt Sustainability Now?

The combination in the past two decades of $r < g$ becoming a pervasive fact across most advanced economies and economists developing a variety or arguments for why the gap between m and r will persist has implications for many economic questions. It has contributed to a rethinking of how the evolution and dynamics of inequality over time, why financial markets misallocate capital, and how the search for safety can trigger economic crises.

For the study of debt sustainability, $r < g$ has meant that the conventional focus of calculating present values of future primary balances became futile: no matter what those balances are forecasted to be, how they respond to debt, or what their maximum is, still their present value is infinity. Fortunately, the classical insights can be rescued by discounting the relevant returns on private capital, as long as $g < m$.

In turn, $r < m$ has meant that flows of debt revenue appear. These revenues have been growing for the last two decades. Population aging has increased the demand for stores of value, the scars of the great financial crisis have increased the demand for safety, and growing financial regulations have increased financial repression while

reducing the offer of private forms of collateral or liquidity. Because these are all structural changes, this has emboldened economists to speculate that the premium on government bonds will persist on average, and so debt revenue can play a significant role in sustaining the public debt moving forward. Whether this is so depends on how large the debt revenue is.

Measurement of the Debt Revenue

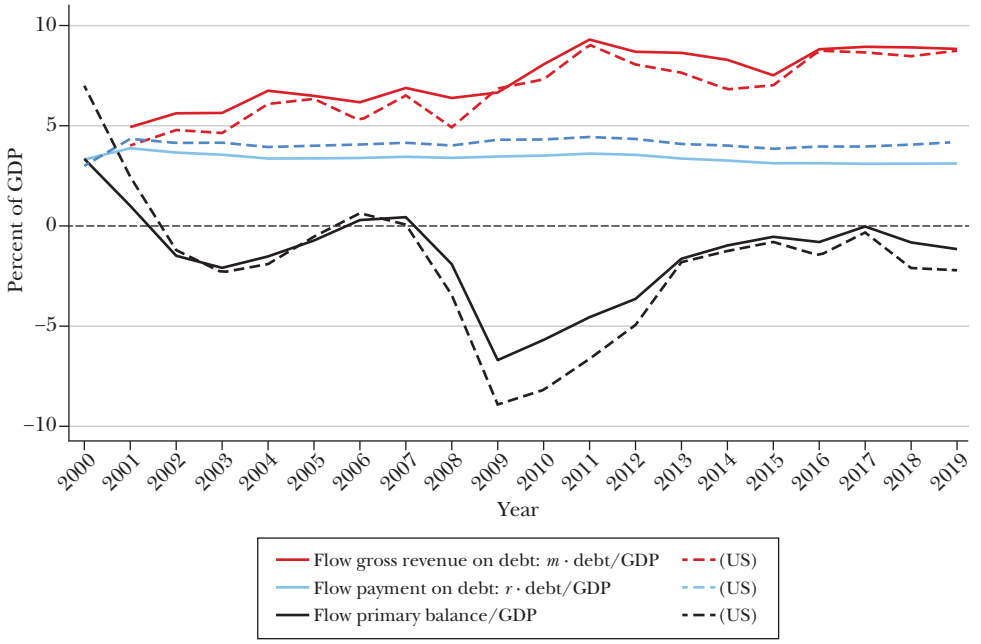
Measuring the debt revenue is hard because it involves measuring the difference between two returns for which there are no immediate counterparts in the data. There are many private investments with different returns, and many ways in which governments borrow (as well as invest). Moreover, it is total returns that matter for the debt revenue, so all of the different returns must be weighted, as opposed to picking just one that is more relevant at the margin.

Some basic calculations give a sense of the likely size of the debt revenue term. Figure 1 shows in one series (in blue) the total interest payments by the G-7 countries—that is, Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States—summed using current exchange rates and divided by total GDP every year, over the last 20 years. Also in the figure, in dashed series, is the series for the United States. This is a direct measure of the return on bonds r multiplied by the debt/GDP ratio. The other series in the figure (in red) comes from multiplying the outstanding public debt over GDP by 0.06 plus average inflation. The choice of 6 percent for real m follows from an approximation that the growth rate of per capita real consumption should equal the difference between the marginal product of capital and the subjective discount rate times the intertemporal elasticity of substitution.[7] Given standard textbook parameter choices like a growth rate of 0.02, a discount rate of 0.02, and an intertemporal elasticity of substitution of 0.5, it follows that $m = 0.06$ plus inflation.

The gap between the two series—the red and the blue in Figure 1—gives the flow revenue on debt: *(m-r)Debt/GDP*. At the start of the twenty-first century, it was around 2 percent of GDP, but by the pre-pandemic years it had climbed above 6 percent, resulting in an average over the 20 years of 3.8 percent. In terms of present values, for $m - g = 0.04$, and ignoring risk or uncertainty in calculating the present value over a long time horizon, debt revenue can sustain 3.8/0.04 = 95 percent of GDP of public debt. For comparison, the 2020 value of debt/GDP for the group of G-7 countries was 140 percent. The debt revenue term over this time was approximately two-thirds of outstanding debt. For the United States during these 20 years, the debt revenue flow was on average 3 percent, for a present value of 75 percent of GDP to compare with the sum of market value of privately held Treasury debt in 2020 (86 percent) and deposits at the Federal

[7]Some readers will recognize this as an application of the Ramsey (1928) formula that specifies an optimal growth rate for consumption.

Figure 1
The Flow Budget Components as a Ratio of GDP for the G-7 Countries and the United States



Source: IMF (1972–2019a, b, 2021b).
Note: Interest payments as a ratio of GDP; public debt as a ratio of GDP times 0.06 plus inflation; primary balance as a ratio of GDP. Adding over all countries in the G-7.

Reserve plus currency in circulation (25 percent). These back-of-the-envelope figures suggest that a significant part of the public debt may be backed by debt revenues.

Another comparison is with the other term in the intertemporal budget constraint for sustainability of sovereign debt, the primary balance. The black series in Figure 1 shows it for the G-7 countries during the last 20 years, again summing across countries and dividing by their total GDP. On average, it has been negative, and smaller in absolute value than the debt revenue. As these rich countries have run large deficits, especially during the great financial crisis and the Covid recession, their outstanding public debt has greatly risen. However, it has risen by significantly less than it would have if the government had paid market interest rates on the new debt, and Figure 1 shows the difference was substantial.

Measuring Debt Revenue as a Residual

In a series of inspiring articles, Jiang et al. (2019; 2020; 2021) measured the expected present value of primary balances using the returns to private investment as the discount rate. They then compared this to the public debt outstanding for

the United States and for many countries in the eurozone. Because the difference between the two in the government budget constraint is the debt revenue term, this provides an approach to measure it as a residual.

To do so, one must have measures of expected future balances and measures of the returns to private investment. For the first, Jiang et al. (2019) use past behavior captured by a regression of US surpluses and other fiscal and macroeconomic variables on their past annual values between 1947 and 2019 as well as CBO estimates of what future deficits will be. For the second, they use an empirical asset-pricing model that can fit the observed returns on stocks and US Treasury bonds within this sample. Their results are puzzling: the debt revenue term is 246 percent of GDP on average, a very large number.

There are three reasons for these extreme estimates. First, since the US primary balance has been, on average, negative for the past seven decades, the present value of this average is negative. Second, this primary balance is strongly procyclical, as governments run deficits during recessions. This makes it a risky flow that is low when money is more valuable, which pushes down its present value when adjusting for risk. Third, because government spending and revenues in the long run move closely with GDP, they carry the long-run risk that seems to drive much of the riskiness in stocks and which leads to their high average returns. This large risk in money terms implies a large negative present value.

One can argue with the precise estimates, and the financial valuation of distant payoffs is as much art as it is science. Moreover, this calculation assumes that debt was sustainable: it takes as given that its market value is high and will remain so and uses this to infer what must be the present value of debt revenue that makes it so. Looking forward, perhaps the United States will suddenly start running large primary surpluses, and maybe these will be higher in future recessions (a terrible idea as procyclical deficits would likely exacerbate the amplitude of business cycles). But unlike what happened after the World Wars, there are no reasonable projections that there will be surpluses even in the distant future. Given the large stock of outstanding debt, this inescapably implies that the debt revenue term must be large.

Looking at other countries, Olijslagers, Van Wijnbergen, and de Vette (2020) focus on the Netherlands, which has often had primary surpluses that are less aggressively procyclical. They find that 53 percent of the outstanding public debt in 2018 is accounted for by the debt revenue term. For the countries in the eurozone, Jiang et al. (2021) find that the variation in the relative convenience yields explains most of the variation in sovereign yields across different countries. They estimate conservatively that since the start of the euro two decades ago, Spain and Italy have earned between 5 percent and 8 percent of GDP less than Germany in debt revenue.

Measuring Debt Revenue by Measuring the Premium

An alternative, more direct approach to measure debt revenue is to measure the premium $m - r$ and multiply it by debt/GDP. The difficulty with doing so concerns how to measure the returns on private capital. There are thousands of alternative investment projects and financial assets in an economy.

One approach provides some sensible estimates. From a macroeconomic perspective, it focuses on the average returns on the aggregate capital stock. From a financial perspective, it follows the teachings of the justly celebrated Modiglian-Miller theorem, looking at the income the project generates while ignoring the way this is carved up across the different financial instruments that funded the projects. Combining the two perspectives suggests dividing the total income that goes to the owners of the capital and firms by the total capital stock in the economy (Reis 2022b).

Table 1 shows a baseline estimate of m for the United States of 8.2 percent. This is close to the 6 percent real return assumed at the start of this section, since inflation has averaged 2.1 percent during these two decades. At the same time, reasonable changes in the assumptions used to measure both the numerator and the denominator can affect these estimates. For instance, in the denominator, the measure of the capital stock used was the standard one produced by the Bureau of Economic Analysis, but this may well be an underestimate of m due to undercounting investment in intangibles during this period. In the numerator, the baseline apportions two-thirds of the income of the self-employed to labor, and one-third to capital. Instead counting all of it as capital income, as the national accounts usually do, raises the estimate of m.

Table 1 shows a few more reasonable alternatives starting from the baseline. Subtracting the corporate taxes that firms pay is straightforward. A more controversial adjustment is whether to subtract rent payments, because land is fixed and is not a capital that the economy can accumulate. At the same time, if these are subtracted, then the increase in the price of the capital stock should perhaps be included as this is a gain to its holder. Across the alternatives, an m between 6.2 percent and 10.5 percent is reasonable, with the baseline estimate roughly in the middle.

The next panel in Table 1 turns to financial markets as a source of data on returns instead. A broad index of stocks is captured is captured by the Wilshire 5000 index, which has between 4000 and 5000 publicly traded firms depending on the year. Over these two decades, US firms increasingly turned to corporate bonds with an expansion of credit flowing through bonds that were rated as being especially risky in terms of default (a credit rating of BBB). These two measures suggest a return between 6.7 percent and 7 percent. However, this is likely an underestimate of m as many firms do not publicly trade stocks or do not issue corporate bonds. In the other direction, focusing on a narrower set of firms that sell shares (those in the S&P 500 index) or on bonds that are less likely to default (those with a AAA credit rating), the estimates are smaller. Most households do not directly hold stocks (although they may hold stocks indirectly via pension, life insurance, or retirement accounts), but many invest in housing, so the table also reports returns on housing that include both the service (or rent) flows from homeownership as well as capital gains and losses. In the other direction, banks actively buy and sell government bonds looking at the alterative returns they would get by lending to other banks at the interbank rate.

Even measuring r is not as easy as it may seem. At the margin, if the US government wants to borrow an extra dollar for one year, then the cost is given

Table 1

Average Nominal Annual Returns (2000–20) in the United States for Measures of *m* and *r*

Measure		%
Return on private investment/Marginal Product of Capital (*m*)		
Income Measure		
(i)	Ratio of Payments to Capital and the Capital Stock	8.2
	(i–a) with adjustment for intangible capital formation	8.0
	(i–b) including proprietors' labor income	10.5
(ii)	(i) minus corporate taxes	7.4
(iii)	(ii) minus rent payments	6.2
(iv)	(iii) plus capital gains	7.1
Financial Measure		
(v)	Wilshire 5000 stock market index	7.0
	(v–a) S&P 500 stock market index	6.6
(vi)	BBB-rated bonds	6.7
	(vi–a) AAA-rated bonds	5.9
(v)	Housing	8.2
(vi)	Interbank rate	2.2
Return on government bonds (r)		
(i)	Return on Treasuries of average maturity	4.1
(ii)	Yield on 1-year Treasuries	1.6

Source: Bureau of Economic Analysis (1901–2020, 1925–2020a, b, c, d, 1929–2020a, b); Global Financial Data (1871–2020); FRED (1960–2020a, b, 1962–2020a, b, 1970–2020, 1986–2020); Jorda et. al (2019).
Note: For detailed description of the series and data sources, see Reis (2022a).

by the interest rate on one-year Treasury bonds. However, the average maturity of government bonds during this period was closer to five years. These bonds gave a significantly higher return on average every year to their holders.

Looking at the whole table, one could make a case for a premium that lies anywhere between 0 and 8.9 percent. More research is needed to pin this down more precisely. The initial estimate at the start of this section set a real *m* of 6 percent and used interest payments, which when divided by the stock of debt for the G-7, leads to an estimate of 1.8 percent for the average real *r*. With a premium of 4.2 percent, and the International Monetary Fund forecasting a net US public debt/GDP ratio between 2021 and 2025 of 103 percent (IMF 2021a), then debt revenue seems likely to play a major role in the sustainability of this debt.

Policy Tradeoffs and Principles of Fiscal Policy

Each of the four sources of the premium on returns that gives rise to debt revenue—store of value, safety, repression, and liquidity—leads to new policy

tradeoffs. Moreover, because some of these policies also affect primary balances, they have the potential to reinforce some of the principles of fiscal management that grew out of the traditional approach, while upending some others.

More Public Debt Is Even Less Sustainable Debt

If there is demand for public debt as a provider of store of value, safety, collateral or liquidity, then increasing its supply should reduce the premium on returns. That premium arises because public debt was scarce. More public debt makes it less special, so it comes with an increase in the returns on public debt and a smaller gap between private investment and public debt. The debt revenue shrinks. Therefore, if the government runs a primary deficit, this not only reduces the classic term of debt sustainability, but also the second term on debt revenue.

Less Austerity: Deficits Can Stimulate the Private Economy

Spending more or cutting taxes during a recession will lower primary balances. However, in classical analyses, this may also stimulate economic activity, which will raise tax revenues and offset some of the decline in the primary balance. With a debt revenue, the temporary increase in the public debt that results will provide the economy with more stores of value, collateral, liquidity, or safety. If these are useful for economic activity or for investment, then this may provide a further stimulus to output. Related to this, public investment may increase the profitability of existing private capital stock, infrastructure being a case in point. Then, the deficits to pay for this investment may raise the returns on private capital, increase the premium, and so partly pay for themselves through debt revenue.

More Austerity: Extraordinary Debts Should Be Paid down Faster

Classic analyses of primary surpluses prescribe that a sudden unexpected increase in public spending, like what happened in 2020 in response to the pandemic, should come with only slightly higher tax revenues. This is because tax rates should be smooth over time in order to minimize their distortions. Primary balances should therefore fall when the spending rises, and then be slightly higher than before in the years that follow to slowly pay down the debt that resulted. From the perspective of debt revenue though, the increase in public debt makes the specialness of public debt less scarce. Weighing this effect, the fiscal authorities may want to raise taxes more aggressively in the short run in order to repay the debt faster. This way, they can enjoy more debt revenue in the future and deliver lower taxes in the long run.

Similarly, beyond stimulating output, there is a case for primary balances to fall during a recession because tax rates are kept unchanged so tax revenue falls. However, the debt revenue may move in the same or opposite direction, depending on whether the shocks that caused it raise or lower the demand for collateral and liquidity. Tax cuts and government spending may satisfy this demand in different directions.

Public Debt Crowds Out and In Private Debt

Savings equal investment. Therefore, for a fixed stock of private savings, if the government saves less by having a deficit, then private investment must fall. Public debt crowds out private investment from the perspective of classical analyses. From the perspective of the specialness of public debt, there are other forms of crowding out and crowding in. For instance, if public debt increases the supply of collateral, it may allow for private savings to rise, increasing investment. Still from the perspective of collateral, private assets that can serve as collateral must sometimes be produced by the private sector. If the supply of public debt crowds out this production, then this serves as a countervailing force on investment.

More subtle, if the government adopts austerity policies, and there are fewer public bonds as a result, then investors will look for which private assets are safe enough to serve as collateral. This makes these private assets more sensitive to information and so less suitable as a whole to serve as collateral. This multiplies the initial effect of austerity in making collateral scarcer in the economy and increasing the premium on returns. It may also trigger a financial crisis due to the absence of collateral.

Debt Management Creates Risks

Traditional debt sustainability analyses emphasize how the response of primary balances to public debt affects the present value of primary balances. However, their responsiveness to debt, output, inflation, or other variables also affects the riskiness of government debt. Therefore, the fiscal response functions also determine the specialness of debt in providing safety, and so the size of debt revenue.

Moreover, say that the government reduces financial repression that made public bonds special, and so lowers debt revenue. To keep public debt sustainable, it offsets this by increasing taxes and so the present value of primary balances. Repression through the efficiency costs of taxation is higher. In addition, with a higher average tax level, future changes in government spending and revenues that cause changes in tax rates may create more uncertainty in returns in the economy and lower investment and economic activity.

Price Stability Keeps Debt Sustainable

Public debt carries a risk of inflation because it promises a fixed nominal payment. Many private investments instead have returns that rise in nominal terms with inflation. Therefore, more inflation risk reduces the premium and the debt revenue. When the public debt is high, it may be more tempting to let inflation rise, temporarily giving debtholders negative returns, as happened in 2022. But it is the trust by investors that monetary policy will do its best to prevent this from happening that allows for the debt revenue term to remain large. Independent inflation-targeting central banks may be especially in the interest of the fiscal authority because price stability—as opposed to attempts to inflate the debt—maximize debt revenue and may keep debt sustainable.

Richer Monetary-Fiscal Interactions

Quantitative easing policies consist of paying for government bonds in exchange for deposits at the central bank. These have different premia, so they come with different debt revenues, which are now partly earned by the central bank and then rebated to the government. This adds a fiscal dimension to monetary policy.

An important difference is that the liabilities of the central bank are the unit of account in the economy—"money," for short. Treasury bonds instead have a price that is set at auction when they are sold and that fluctuates in markets. Therefore, while the market value of debt can quickly adjust to shocks to primary surpluses or to debt revenue, the real value of money only adjusts slowly with changes in the prices of goods. Debt sustainability is tied to price stability (Calvo and Velasco 2022).

Finally, imagine that monetary policy keeps nominal interest rates fixed. This could be by choice, or it could be because the central bank would like to lower nominal interest rates but they have reached an effective lower bound. If inflation is sticky, then traditional analyses note that more government spending can stimulate output and so increase primary balances. Because issuing more debt now has no impact on the real return r, it also raises debt revenues (Mian, Straub, and Sufi 2022).

Spillovers across Borders

The debt of the US government is seen as a safe haven by international investors, and this is a significant part of the debt revenue that it collects. Other countries never have debt revenue that is too large and, worse, any existing debt revenue in those countries can dissipate quickly during a financial crisis when investors rush out of all domestic assets. At all times, this means that the fiscal (and monetary policy) of the United States will spill over to the returns premium of countries around the world and affect their debt revenue and debt sustainability (Jiang, Krishnamurthy, and Lustig 2020).

Debt Revenue and Ricardian Equivalence

Imagine that the government provides a transfer to a household, funds it by selling a bond to that household, and later on pays for that bond by taxing the same household again. The principle of Ricardian equivalence states that the household will save the whole of the initial transfer in order to pay for the future taxes and change no other of its choices. With a premium on government debt, the household may be willing to collect a low return on the public bond issued by the government to finance the transfer. This is an opportunity cost for the household that could be collecting higher returns on private investment. This cost is just offset by the debt revenue and by lower taxes in the future to pay for the debt. Therefore, the household still realizes its net wealth has not changed and changes none of its other actions (Barro 2020).

What Is Good for the Public Purse May Not Be Good for Welfare

Any improvement in how the private credit market works or in social programs that reduce the supply of savings will reduce the demand for the safety or store of

value that is offered by public debt. Therefore, it lowers debt revenues. Policies that promote financial development, provide social insurance, or lower inequality may be good for economic growth and social welfare, but they also reduce debt revenue and hence shrink the fiscal resources available for other government programs. When considering public policies, governments may want to take into account not only their direct impact on the primary balance, but also how much they will affect the usefulness and demand for public debt

Moreover, just because there are debt revenues does not mean that society would be better off if there was more public debt. After all, if the government can just increase the supply of public debt at no cost, it might want to do so until the demand for the specialness of this debt is fully satiated. At that point, both the return premium and the debt revenue are zero.[8] At the same time, having positive debt revenues can lower the need to use distortionary taxes to collect revenues in response to shocks and be used to stimulate aggregate demand out of deep recessions. More generally, the different policy trade-offs described so far combine to imply an optimal amount of debt.

Conclusion

The traditional literature on debt sustainability has focused on measuring the expected present value of primary balances and on studying how different policies may increase or lower it, depending on the relative strength of different trade-offs. This literature has its challenges, and there are still many open questions both in theory and in measurement, but it has been useful to fiscal authorities all over the world when considering how much spending and how much borrowing a government can do. However, the steady downward trend in the return on government bonds, which for years leading up to the pandemic was decidedly below the growth rate of the economy, has made the framework hard to apply because the present value of future primary surpluses is not mathematically well defined.

At the same time, the returns to private investment in the data have stayed comfortably above the growth rate of the economy, and there has been a wealth of theories to explain why there is an increasing discount in government bond returns relative to private investment. Taking into account this premium on government bond returns reveals a new fiscal revenue that comes from the act of issuing public debt to satisfy the demand for its store of value, safety, collateral, or liquidity. Simple calculations suggest that this debt revenue term is large and may be sustaining most of the public debt outstanding in developed economies. Perhaps this accounts for the lack of a debt crisis in the United States and most other advanced economies in spite of debt/GDP ratios that are broaching record highs.

[8]This argument is an extended version of the famous Friedman (1969) rule for the supply of currency, which held that the optimal quantity of money should be so that the level of price deflation in the economy would cause the nominal interest rate to be zero and the seignorage to be nil.

Economists around the world are debating the path of deficits and debt. For example, US economists are discussing how quickly to pay for the pandemic debt; European Union economists are considering what rules might be useful for restricting national government deficits and debt; and economists who study emerging and low-income economies are debating whether a sovereign debt crisis is on the horizon. For all of these debates, and many others, considering the debt revenue term promises to be useful.

■ *I am grateful to Erik Hurst for the encouragement to write this essay, to Erik Hurst, Nina Pavcnick, and Timothy Taylor for comments, to Marina Feliciano and Borui Niklas Zhu for research assistance, and to audiences at conferences hosted by the Banque de France / CEPR and the European Central Bank for their comments. The article did not receive any financial support from any source.*

References

Abbas, S. Ali, Alex Pienkowski, and Kenneth Rogoff. 2020. *Sovereign Debt: A Guide for Economists and Practitioners*. Oxford: Oxford University Press.

Alesina, Alberto, Carlo Favero, and Francesco Giavazzi. 2019. *Austerity: When it Works and When it Doesn't*. Princeton, NJ: Princeton University Press.

Angeletos, George-Marios, Fabrice Collard, and Harris Dellas. 2016. "Public Debt as Private Liquidity: Optimal Policy." CEPR Working Paper 15488.

Barro, Robert J. 2020. "R minus G." NBER Working Paper 28002.

Bassetto, Marco, and Wei Cui. 2018. "The Fiscal Theory of the Price Level in a World of Low Interest Rates." *Journal of Economic Dynamics and Control* 89: 5–22.

Bassetto, Marco, and Wei Cui. 2021. "A Ramsey Theory of Financial Distortions." IFS Working Paper W21/05.

Bayer, Christian, Benjamin Born, and Ralph Luetticke. 2020. "The Liquidity Channel of Fiscal Policy." CEPR Discussion Paper 14883.

Berentsen, Aleksander, and Christopher Waller. 2018. "Liquidity Premiums on Government Debt and the Fiscal Theory of the Price Level." *Journal of Economic Dynamics and Control* 89 (C): 173–82.

Blanchard, Olivier. 2019. "Public Debt and Low Interest Rates." *American Economic Review* 109 (4): 1197–1229.

Bohn, Henning. 1998. "The Behavior of US Public Debt and Deficits." *Quarterly Journal of Economics* 113 (3): 949–63.

Bonam, Dennis. 2021. "A Convenient Truth: The Convenience Yield, Low Interest Rates and Implications for Fiscal Policy." De Nederlandsche Bank Working Paper 700.

Brunnermeier, Markus K., Sebastian A. Merkel, and Yuliy Sannikov. 2022. "Debt as a Safe Asset." NBER Working Paper 29626.

Bureau of Economic Analysis. 1901–2020. "Table 2.7. Investment in Private Fixed Assets, Equipment, Structures, and Intellectual Property Products by Type." United States Department of Commerce. https://apps.bea.gov/iTable/iTable.cfm?reqid=10&step=3&isuri=1&table_list=51&series=q&first_year=2013&allyears=1&tabledisplay=&scale=-99&last_year=2020 (accessed September 6, 2022).

Bureau of Economic Analysis. 1901–2020. "Table 2.7. Investment in Private Fixed Assets, Equipment, Structures, and Intellectual Property Products by Type." United States Department of Commerce. https://apps.bea.gov/iTable/iTable.cfm?reqid=10&step=3&isuri=1&table_list=51&series=q&first_year=2013&allyears=1&tabledisplay=&scale=-99&last_year=2020 (accessed September 6, 2022).

Bureau of Economic Analysis. 1925-2020a. "Table 6.1. Current-Cost Net Stock of Private Fixed Assets by Industry Group and Legal Form of Organization." United States Department of Commerce. https://apps.bea.gov/iTable/iTable.cfm?reqid=10&step=3&isuri=1&table_list=41&series=q&first_year=2013&allyears=1&tabledisplay=&scale=-99&last_year=2020 (accessed September 6, 2022).

Bureau of Economic Analysis. 1925–2020b. "Table 6.2. Chain-type Quantity Indexes for Net Stock of Private Fixed Assets by Industry Group and Legal Form of Organization." United States Department of Commerce. https://apps.bea.gov/iTable/iTable.cfm?reqid=10&step=3&isuri=1&table_list=48&series=q&first_year=2013&allyears=1&tabledisplay=&scale=-99&last_year=2020 (accessed September 6, 2022).

Bureau of Economic Analysis. 1925–2020c. "Table 1.1. Current-Cost Net Stock of Fixed Assets and Consumer Durable Goods." United States Department of Commerce. https://apps.bea.gov/iTable/iTable.cfm?reqid=10&step=3&isuri=1&table_list=16&series=q&first_year=2013&allyears=1&tabledisplay=&scale=-99&last_year=2020 (accessed May 31, 2022).

Bureau of Economic Analysis. 1925–2020d. "Table 1.3. Current-Cost Depreciation of Fixed Assets and Consumer Durable Goods." United States Department of Commerce. https://apps.bea.gov/iTable/iTable.cfm?reqid=10&step=3&isuri=1&table_list=86&series=q&first_year=2013&allyears=1&tabledisplay=&scale=-99&last_year=2020 (accessed June 15, 2022).

Bureau of Economic Analysis. 1929–2020a. "Table 1.10. Gross Domestic Income by Type of Income." United States Department of Commerce. https://apps.bea.gov/iTable/iTable.cfm?reqid=19&step=3&isuri=1&select_all_years=1&nipa_table_list=51&series=a&first_year=2020&last_year=2021&scale=-9&categories=survey&thetable= (accessed September 6, 2022).

Bureau of Economic Analysis. 1929–2020b. "Table 1.1.4. Price Indexes for Gross Domestic Product." United States Department of Commerce. https://apps.bea.gov/iTable/iTable.cfm?reqid=19&step=3&isuri=1&select_all_years=1&nipa_table_list=4&series=a&first_year=2020&last_year=2021&scale=-99&categories=survey&thetable= (accessed September 6, 2022).

Calvo, Guillermo A. 1988. "Servicing the Public Debt: The Role of Expectations." *American Economic Review* 78 (4): 647–61.

Calvo, Guillermo A., and Andrés Velasco. 2022. "Joined at the Hip: Monetary and Fiscal Policy in a Liquidity-Dependent World." NBER Working Paper 29865.

Cochrane, John H. 2021. "r<g." Unpublished.

Corrado, Carol, Jonathan Haskel, Cecilia Jona-Lasinio, Massimiliano Iommi. 2016. "Intangible Investment in the EU and US before and since the Great Recession and Its Contribution to Productivity Growth" In Investment and Investment Finance in Europe, 73–102. Luxembourg: European Investment Bank.

D'Erasmo, Pablo, Enrique Mendoza, and Jiang Zhang. 2016. "What is a Sustainable Public Debt?" In *Handbook of Macroeconomics*, Vol. 2, edited by J. B. Taylor and Harald Uhlig, 2493–2597. Amsterdam: Elsevier.

Eichengreen, Barry, Asmaa El-Ganainy, Rui Esteves, and Kris J. Mitchener. 2021. *In Defense of Public Debt*, Oxford: Oxford University Press.

Elenev, Vadim, Tim Landvoigt, Patrick J. Schultz, and Stijn Van Nieuwerburgh. 2021. "Can Monetary Policy Create Fiscal Capacity?" NBER Working Paper 29129.

FRED. 1960–2020a. "ICE Data Indices: Bank of America Merrill Lynch US Corp BBB Total Return Index." Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/BAMLCC0A4BBBTRIV (accessed September 7, 2022).

FRED. 1960–2020b. "ICE Data Indices: Bank of America Merrill Lynch US Corp AAA Total Return Index." Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/BAMLCC0A1AAATRIV (accessed September 7, 2022).

FRED. 1962–2020a. "Market Yield on U.S. Treasury Securities at 1-Year Constant Maturity, Quoted on an Investment Basis (DGS1)." St. Louis: Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DGS1 (accessed December 13, 2021).

FRED. 1962–2020b. "Market Yield on U.S. Treasury Securities at 5-Year Constant Maturity, Quoted on an Investment Basis (DGS5)." Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DGS5 (accessed December 13, 2021).

FRED. 1970–2020. "Wilshire 5000 Total Market Full Cap Index (WILL5000INDFC)." Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/WILL5000INDFC (accessed September 9, 2022).

FRED. 1986–2020. "ICE Benchmark Administration: 12-Month London Interbank Offered Rate." Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/USD12MD156N (accessed December 13, 2021).

Friedman, Milton. 1969. *The Optimal Quantity of Money and Other Essays.* New York, NY" Routledge.

Gale, William G. 2019. "Fiscal Policy with High Debt and Low Interest Rates." In *Maintaining the Strength of American Capitalism*, edited by Melissa S. Kearney and Amy Ganz, 78–115. Washington, DC: Aspen Institute Economic Strategy Group.

Gersbach, Hans, Jean-Charles Rochet, and Ernst-Ludwig von Thadden. 2022. "Public Debt and the Balance Sheet of the Private Sector." CEPR Discussion Paper 17529.

Global Financial Data. 1871–2020. "S&P 500 Total Return Index." S&P Dow Jones Indices. (accessed September 7, 2022)

Gorton, Gary, and Guillermo Ordoñez. 2022. "The Supply and Demand for Safe Assets." *Journal of Monetary Economics* 125: 132–47.

Hilscher, Jens, Alon Raviv, and Ricardo Reis. 2022. "Inflating Away the Public Debt? An Empirical Assessment." *Review of Financial Studies* 35 (3): 1553–95.

IMF. 2021a. *Fiscal Monitor: October 2021.* Washington, DC*:* International Monetary Fund.

IMF. 2021b. "World Economic Outlook Database: October 2021." World Economic and Financial Surveys. https://www.imf.org/en/Publications/WEO/weo-database/2021/October/download-entire-database (accessed December 13, 2021).

INTAN-Invest. 1977–2017. "Unpublished Update to Corrado and Hulten (2010) for INTAN-Invest and the SPINTAN project." http://www.intaninvest.net/charts-and-tables/ (accessed June 16, 2022).

Jiang, Zhengyang, Arvind Krishnamurthy, and Hanno Lustig. 2020. "Dollar Safety and the Global Financial Cycle." NBER Working Paper 27682.

Jiang, Zhengyang, Hanno Lustig, Stijn Van Nieuwerburgh, and Mindy Z. Xiaolan. 2019. "The US Public Debt Valuation Puzzle." CEPR Discussion Paper 16082.

Jiang, Zhengyang, Hanno Lustig, Stijn Van Nieuwerburgh, and Mindy Z. Xiaolan. 2020. "Manufacturing Risk-Free Government Debt." CEPR Discussion Paper 16304.

Jiang, Zhengyang, Hanno Lustig, Stijn Van Nieuwerburgh, and Mindy Z. Xiaolan. 2021. "Bond Convenience Yields in the Eurozone Currency Union." Unpublished.

Jiang, Wei, Thomas J. Sargent, Neng Wang, and Jinqiang Yang. 2022. "A p Theory of Government Debt and Taxes." NBER Working Paper 29931.

Jorda, Oscar, Katharina Knoll, Dmitry Kuvshinov, Moritz Schularick, Alan M. Taylor. 2019. "Replication Data for: The Rate of Return on Everything, 1870–2015." Harvard Dataverse. https://doi.org/10.7910/DVN/GGDQGJ.

Krishnamurthy, Arvind, and Annette Vissing-Jorgensen. 2012. "The Aggregate Demand for Treasury Debt" *Journal of Political Economy* 120 (2): 233–67.

Mehrotra, Neil R., and Dmitriy Sergeyev. 2021. "Debt Sustainability in a Low Interest Rate World." *Journal of Monetary Economics* 124: S1–18.

Mian, Atif, Ludwig Straub, and Amir Sufi. 2022. "A Goldilocks Theory of Fiscal Policy." NBER Working Paper 29707.

Miao, Jianjun, and Dongling Su. 2021. "Fiscal and Monetary Policy Interactions in a Model with Low Interest Rates." Unpublished.

Müller, Andreas, Kjetil Storesletten, and Fabrizio Zilibotti. 2019. "Sovereign Debt and Structural Reforms." *American Economic Review* 109 (12): 4220–59.

Olijslagers, Stan, Sweder Van Wijnbergen, and Nander de Vette. 2020. "Debt Sustainability when $r - g < 0$: no Free Lunch After All." CEPR Discussion Paper 15478.

Ramsey, Frank. 1928. "A Mathematical Theory of Saving." *The Economic Journal* 38 (152): 543–59.

Reis, Ricardo. 2021. "The Constraint on Public Debt when $r < g$ but $g < m$." CEPR Discussion Paper 15950.

Reis, Ricardo. 2022a. "Replication data for: Debt Revenue and the Sustainability of Public Debt." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E180241V1.

Reis, Ricardo. 2022b. "Which r-star, Government Bonds or Private Investment? Measurement and Policy Implications." Unpublished.

Schmid, Lukas, Yang Liu, and Amir Yaron. 2021. "The Risks of Safe Assets." CEPR Discussion Paper 16407.

Sims, Christopher. 2020. "Optimal Fiscal and Monetary Policy with Distorting Taxes." Unpublished.

Willems, Tim, and Jeromin Zettelmeyer. 2022. "Sovereign Debt Sustainability and Central Bank Credibility." *Annual Review of Financial Economics* 14.

Yared, Pierre. 2019. "Rising Government Debt: Causes and Solutions for a Decades-Old Trend." *Journal of Economic Perspectives* 33 (2): 115–40.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 125–146

Fiscal Histories

John H. Cochrane

What fundamentally drives inflation or deflation, or the value of money? The *fiscal theory of the price level* offers a novel answer to this age-old question. It is particularly relevant today, with inflation that seems related to large recent deficits, and given the foundational difficulties widely acknowledged in traditional monetary, Keynesian, and New Keynesian theories.

In this essay, I offer narrative discussions of how fiscal theory can account for prominent episodes when inflation did, or did not, occur. Why did inflation rise in the 1970s and fall in the 1980s? Why was inflation quiet in the 2010s, but then rose in 2021? Why does inflation fall in recessions and rise in booms? These stories help us to see how fiscal theory works and how to apply it in practice, more transparently than by staring at forests of equations.

The fact that there are such plausible stories—that fiscal theory can plausibly account for historical episodes—is news, since many economists and commentators seem to think that fiscal theory can be quickly dismissed by well-known episodes. Formal quantitative analysis and evaluation builds on plausible stories, and I hope this essay and my recent book (Cochrane forthcoming) inspire additional formal analysis.

■ *John H. Cochrane is the Rose-Marie and Jack Anderson Senior Fellow at the Hoover Institution at Stanford University, Stanford, California. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. His email address is john.cochrane@stanford.edu and his webpage is https://www.johnhcochrane.com.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.125.

Fiscal Theory

First, I briefly describe fiscal theory and how it contrasts with conventional theories of inflation. The fiscal theory states that *inflation adjusts so that the real value of government debt equals the present value of primary surpluses.*

Most simply, money is valuable because we need money to pay taxes. If, on average, people have more money than they need to pay taxes, they try to buy things, driving up prices. In the words of Adam Smith (1776 [1930], Book II, chap. II): "A prince, who should enact that a certain proportion of his taxes be paid in a paper money of a certain kind, might thereby give a certain value to this paper money . . ." Taxes are a percentage of income. Thus, as prices and wages rise, your dollar income rises, and the amount of money you must pay in taxes rises. A higher price level soaks up excess money with tax payments. Equivalently, the real value of money, the amount of goods and services a dollar buys, declines as the price level rises. But the real value of taxes does not change (much), so a higher price level lowers the real value of money until it equals the real value of tax payments.

This story is simplistic. We add more realistic ingredients in order to make the theory useful to think about economic events and policy.

First, the government also spends and transfers money to people. So money is soaked up by government *surpluses*, the excess of taxes over spending, not just by taxes.

Second, governments also sell bonds. If you buy a one-year bond, you give the government $1 today, and the government gives you $1 plus interest, say $1.05, in a year. So, the government must print up money to redeem bonds that come due, which pushes toward inflation. But the government can also soak up money by selling new bonds. The government can run deficits, a negative surplus, by selling bonds. But the government cannot keep rolling over its debts forever, issuing new bonds to repay old bonds. Eventually, all of the money outstanding today and all of the money promised by outstanding government debt must be soaked up by surpluses. Thus, prices adjust until the real value of *all government debt*, including money, equals the present value of current *and future* surpluses.[1]

The economic logic is the same as the basic way we think of stock and bond prices. The stock or bond price adjusts so that the value of a stock or bond is equal to the expected discounted present value of dividends or coupons. Government bonds, repaid by surpluses, are effectively invested in the government.

[1] In equations, the price level adjusts so that

$$\frac{B_{t-1}}{P_t} = E_t \sum_{j=0}^{\infty} \frac{\Lambda_{t+j}}{\Lambda_t} s_{t+j}$$

where B = nominal government debt, P = price level, Λ is a stochastic discount factor such as marginal utility or the inverse cumulative return, and s is the real primary government surplus. Money (cash and reserves) are part of B.

The insight that the *present value* of surpluses matters for inflation quickly surmounts some armchair rejections of fiscal theory and opens the door to a more interesting interpretation of events. One might think fiscal theory predicts a strong correlation between debt or deficits and inflation. Yet while inflation is sometimes linked to debt and deficits, there is often little or even lower inflation with deficits, as in many recessions, and there is often inflation without big deficits. However, fiscal theory does *not* necessarily predict a tight relationship between current debt or deficits and inflation. If the government runs a big deficit, but people trust that deficit will be repaid by higher subsequent surpluses, then people are happy to hold the extra debt rather than try to spend it, and there is no inflation. That hypothesis is sensible. When a corporation borrows to build a factory, it runs a big deficit, and then slowly pays off the bonds, a long stream of surpluses. Governments that want to borrow, to raise revenue to fight wars or recessions, and do not want to create inflation, will credibly promise repayment. Fiscal theory only predicts inflation when debt is larger than what people think the government will repay.

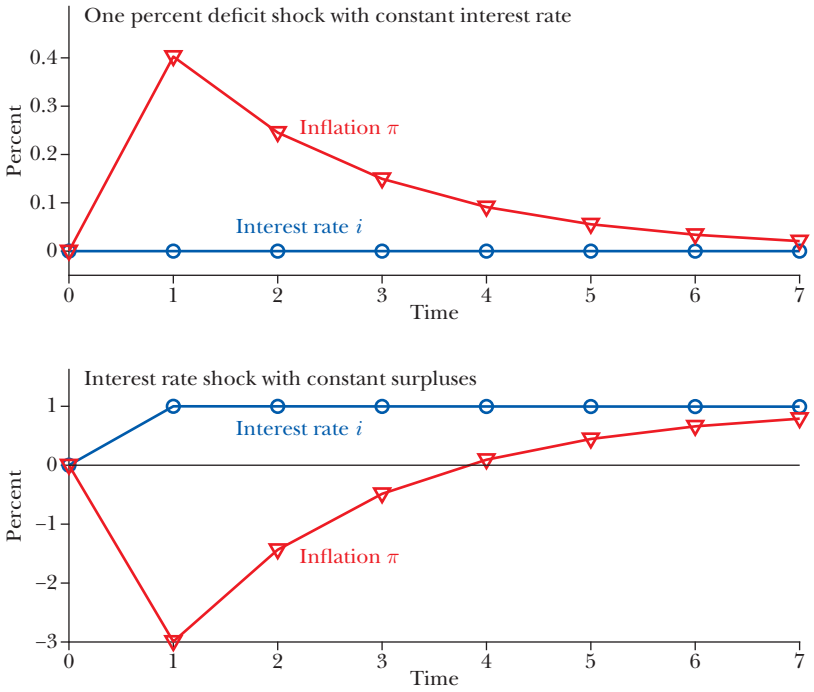
Inflation often seems to come from nowhere or to over-respond to small shocks. Well, news or sentiment about long-run fiscal surpluses can change quickly, as do investor's views of stocks. Moreover, dynamics that resemble bank runs underlie the present value. Many governments roll over a lot of short-term debt. People may dump government debt today, simply because they fear nobody will roll over the debt next year. Iterating forward, we economists see the present value of surpluses, but that long-term view is not necessarily in the minds of bondholders.

Most of all, *discount rates matter* to present values. When interest rates rise, bond values fall. A higher real interest rate makes the same stream of expected surpluses less valuable. So, higher real interest rates lower the value of debt and act as an inflationary force even with no surplus news. Equivalently, a higher real interest rate means that the government has to pay more to finance its debt. As we shall see, this variation in discount rates or interest costs is central to understanding post–World War II US inflation.

In fiscal theory, the central bank can still set a nominal interest rate target, as central banks do. If prices are flexible, the nominal interest rate target determines expected inflation. Central banks remain powerful! However, shocks to the present value of surpluses then determine unexpected inflation, devaluing nominal debt to match the lower present value of surpluses. There is nothing central banks can do to stop that.

Sticky prices produce interesting and realistic dynamics and allow inflation to affect the real economy. The interest rate target then determines the timing of inflation and the long-run level of inflation, while fiscal shocks produce a drawn-out unexpected inflation. Figure 1 presents a simulation of these effects, for concreteness but also for reassurance that fiscal theory really does lend itself to quantitative formal modeling, not just storytelling. I use the simplest standard sticky-price

Figure 1
Inflation Response to Deficit and Interest Rate Shocks



Source: Author's calculations. See Cochrane (2022).
Note: The graphs plot how inflation evolves in a fiscal-theory model with sticky prices. Top: There is a 1 percent deficit shock at time 1, with no change in interest rate. Bottom: There is a one percentage point permanent rise in interest rate, as shown, with no change in surplus or deficit. The model is given in footnote 2.

model.[2] Sticky prices are represented in the Phillips curve: Firms can only change prices infrequently, but do so looking forward. If other prices will be higher next year, firms raise prices now. As a result, inflation today is driven by expected inflation next year, plus additional pressure if demand is high.

[2]The model is

$$
\begin{aligned}
x_t &= E_t x_{t+1} - \sigma(i_t - E_t \pi_{t+1}) \\
\pi_t &= E_t \pi_{t+1} + \kappa x_t \\
i_{t+1} &= \eta i_t + \varepsilon_{i,t+1} \\
\rho v_{t+1} &= v_t + r^n_{t+1} - \pi_{t+1} - \tilde{s}_{t+1} \\
E_t r^n_{t+1} &= i_t \\
r^n_{t+1} &= \omega q_{t+1} - q_t
\end{aligned}
$$

where x_t = output gap, i_t = interest rate, π_t = inflation, v_t = real value of debt, $\tilde{s}_t$ = real primary surplus scaled by the value of debt, r^n_{t+1} = nominal return on government debt, q_t = price of the government debt bond portfolio. $\omega = 0.9$ describes a geometric maturity structure of debt. $\sigma = 0.5$, $\kappa = 0.5$, $\eta = 1$, $\rho = 0.98$.

The top panel simulates what happens to inflation after a 1 percent fiscal shock—the sum of current and expected future deficits rises 1 percent—while the central bank does not change interest rates. The unexpected deficits create a protracted inflation. This is, roughly, what I argue below happened in 2021. Bondholders lose from the long period in which inflation is higher than the interest rate. This is a more realistic prediction than an instantaneous price-level jump.

Indeed, in continuous time the price level does not move at all on the day of the fiscal shock. The entire decline in value of debt comes from this period of drawn-out inflation. It is a fiscal theory of inflation, of the long-run price level, but not of price level jumps, of the initial price level.

The bottom panel presents what happens to inflation and output if the Fed raises interest rates one percentage point and there is no change in fiscal surpluses. By including long-term debt, this simulation expresses the common idea that an interest rate rise temporarily lowers inflation and output. But the higher interest rates eventually raise inflation. In both cases output, not shown, roughly follows inflation.

The two simulations of Figure 1 offer an important benchmark for understanding events and analyzing fiscal and monetary policy. Historical episodes and policies mix fiscal and monetary interventions, and thus mix the two simulations. (Most interest rate hikes die out more quickly, and we do not see the long-run positive effect.)

Money and Aggregate Demand

The most familiar theory of inflation is based on money supply and money demand: Inflation comes from too much money chasing too few goods and services.

At first glance, the monetary and fiscal stories sound similar. And that is a good thing for fiscal theory. If you live in a fiscal theory economy, as I think you do, you wouldn't immediately notice anything unusual just by looking out the window, and neither would the generations of smart economists who have come before us.

But there are crucial differences. First, *which* money is inflationary? In the monetarist story, assets such as checking accounts, created by banks, satisfy money demand, and so are just as inflationary as government-provided cash. Thus, the government must control checking accounts and other "inside" liquid assets. In the basic fiscal theory, only government money, cash and bank reserves, matter for inflation. Your checking account is an asset to you but a liability to the bank, so more checking accounts do not make the private sector as a whole feel wealthier and desire to spend more. The government need not control the quantity of checking accounts and other liquid assets. However, in the basic fiscal theory, government debt, which promises money, is just as inflationary as money itself. Reserves and cash are just overnight government debt.[3]

[3]Reserves are accounts that banks hold at the Federal Reserve. Banks may freely convert reserves to cash and back. The Fed issues cash and reserves, and invests in Treasury debt, just like a giant money-market fund. Because the interest the Fed pays on reserves comes from the interest it gets from Treasury

What about episodes in which we see inflation or hyperinflation clearly caused by printing money? In these episodes, governments print money to finance intractable *fiscal* deficits. They are expansions of government debt relative to the government's ability to repay debt. They are equally inflationary in fiscal theory. Similarly, Milton Friedman once joked that the government could easily cause inflation by dropping money from helicopters. But dropping money from helicopters is a *fiscal* operation, a transfer payment.

The key question is whether *exchanging* money for debt causes inflation. If the central bank issues reserves or cash but takes government debt in return, does *that* inflate? This "open market operation" is what central banks do. In the monetarist view, yes. In the basic fiscal view, no. People and banks really do not care much about holding Treasury debt directly versus holding interest-paying reserves that are backed by Treasury debt. It's like taking your $20 bills and giving you two $5s and a $10.

The monetary theory isn't wrong. It just doesn't apply to today's economy. First, monetary theory requires a meaningful distinction between "money," special assets used for transactions and "liquidity" purposes, and "bonds" or savings vehicles, and that money pays substantially less interest than bonds. This precondition is rapidly evaporating. Second, and more importantly, monetary theory requires that the government controls the money supply. But our central banks do not begin to control the supply of money (like M1 or M2) or liquid assets. The Federal Reserve eliminated reserve requirements altogether in 2020. Instead, central banks set interest rate targets.

By contrast, fiscal theory is consistent with uncontrolled inside money, financial and payments innovation, cryptocurrency, interest rate targets, unstable money demand, elastic money supply policies, and the disappearance of a meaningful distinction between monetary and investment assets, all of which vitiate monetary theory. All that matters, to first order, is total government liabilities—Treasury debt, cash, and reserves—relative to expected repayment.

A Keynesian, looking out the window at a fiscal inflation, sees too much aggregate demand. Fiscal theory agrees, but gives a deeper source of that demand: People can only demand more of all goods, services, and private assets by demanding less government debt.

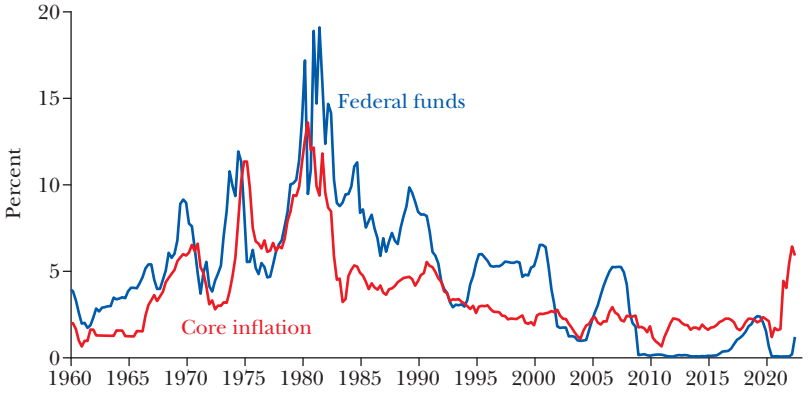
US Inflation History

I take a narrative tour of recent US inflation history to see how events can be interpreted via fiscal theory.

---

securities, and since it remits any profits to the Treasury, we really can unite Fed and Treasury balance sheets and consider cash and reserves as very short-term and liquid forms of government debt, at least to first order.

Figure 2
Core Consumer Price Index Inflation and Federal Funds Interest Rate



Source: Board of Governors (2022), US Bureau of Labor Statistics (2022a).

The Great Inflation

The United States last experienced a major inflation in the 1970s, which ended swiftly in the early 1980s. Figure 2 presents inflation as measured by the core (less food and energy) Consumer Price Index and the federal funds interest rate.[4] Inflation came in three great waves during this time, punctuated by recessions.

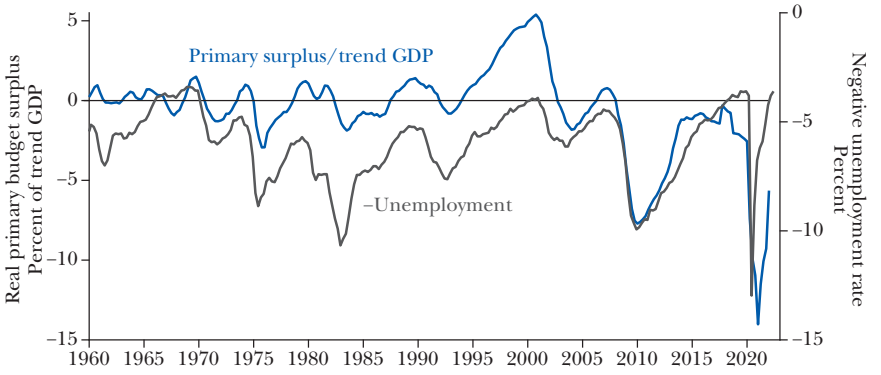
The conventional story of this episode focuses on monetary policy: Monetary policy was too loose in the 1960s and 1970s, accommodated the oil price shocks of the 1970s, and reacted too slowly to inflation. Inflation was conquered in the early 1980s by persistently high real interest rates, at the cost of two bruising recessions.

The fiscal side of the 1970s and 1980s is less well investigated, but suggestive. Figure 3 plots the real primary surplus.[5] The graph emphasizes that most variation in deficits and surpluses comes from business cycles: Tax receipts fall in a recession, and spending on social programs and discretionary stimulus rises. Surpluses rise in the subsequent expansion. We have to see changes in the present value of surpluses on top of this regular pattern. I plot the inverse unemployment rate to allow some eyeball correction for business cycles.

[4]Federal funds are overnight unsecured loans between banks, borrowing and lending reserves. The federal funds interest rate is what the Federal Reserve targets most directly. Core inflation is less volatile than the full CPI.

[5]I plot the primary budget surplus, not including interest costs on the debt, as the value of government debt equals the present value of real primary surpluses. Interest costs enter the government debt valuation equation through the discount rate. I scale by a linear trend fit to log GDP, rather than GDP itself, as we want to see how greater surpluses induced by greater GDP help to repay debts. I plot quantities from the national income and product accounts for easy reproducibility, but they are not ideal measures, especially of interest costs of the debt.

Figure 3
Real Primary Budget Surplus and Negative Unemployment Rate



Source: US Bureau of Economic Analysis (2022a, b, c); US Bureau of Labor Statistics (2022b); author's calculation in Cochrane (2022b).
Note: Real primary surplus is federal taxes minus spending, not including interest costs on the debt. It is measured as federal net lending or borrowing plus federal interest payments, converted to 2012 dollars via the GDP deflator, and divided by trend real GDP, fit by a linear regression to the log of real GDP.

Standard economic history recognizes fiscal deficits of the late 1960s as a cause for inflation, attributing this to President Lyndon B. Johnson's desire to fund both the Great Society and the Vietnam War without tax increases. However, these deficits are smaller than deficits that came later, and did not cause inflation.

The stresses of the Bretton Woods era are less well studied and may help to explain why relatively small deficits caused inflation. Under Bretton Woods, the United States committed to exchange dollars for gold at a fixed rate with foreign central banks. This system, effectively tying the price *level* to gold, is incompatible with any sustained inflation, even the few percentage points of the 1960s. Moreover, international capital markets were largely closed, so the US could not finance trade or fiscal deficits by selling assets abroad. It had to finance trade deficits with dollars—and thereby gold. These constraints provoked essentially a foreign exchange and debt crisis in 1971, when the United States fully abandoned the gold and currency pegs of Bretton Woods, and thereby abandoned its commitment to running fiscal policy consistent with those pegs. Bretton Woods amounted to a precommitment against persistent trade deficits and foreign financing of fiscal deficits, which the United States now practices on a grand scale. Bretton Woods made US currency more fragile when it wished to violate those commitments.

The US economy suffered a slowdown in the 1970s and a break in fiscal policy. Since World War II, the US government had run moderate deficits in recessions, steady primary surpluses in expansions, and slowly reduced the debt/GDP ratio. The 1970s growth slowdown and severe recessions provoked much larger deficits,

lower surpluses in expansions, and less GDP in the debt/GDP denominator. The 1975 primary deficit in particular was larger than any since World War II. Together with the abandonment of Bretton Woods, concern rose over long-run US debt sustainability.

One can digest the waves of inflation in the 1970s with the two experiments of Figure 1. Fiscal shocks lead to sustained inflation. Monetary policy, tightening interest rates without any change in fiscal policy, alleviates inflation temporarily but eventually makes inflation worse. Sims (2011) offers this analysis and calls that pattern "stepping on a rake."

Inflation was not conquered in the 1980s by monetary policy alone. First, although there was much commentary about "Reagan deficits" in the early 1980s, fiscal policy was not in fact tremendously loose. *Primary* deficits, reflecting tax and spending decisions, were unremarkable, especially given the severe recessions of 1980 and 1981–1982. A big part of the measured deficit was simply higher interest costs on existing debt. Second, economic reforms of the time likely encouraged economic growth and thereby helped the budget. Deregulation, starting with airlines, trucks, and telephones, began under President Jimmy Carter, and it, or simply a halt to the regulatory trend of the 1970s, continued in the 1980s. Tax reforms in 1982 and 1986 improved incentives by slashing federal marginal tax rates—the top personal marginal tax rate fell from 70 percent to 28 percent—while removing many exclusions and "loopholes," and thereby broadening the tax base. Whether for these reasons or just good luck, economic growth rose, tax revenues rose, and so did surpluses. As Figure 3 shows, by the late 1990s, the government was running large primary surpluses, fully justifying the 1980s disinflation. The actual surpluses were even larger than shown in the figure, as trend real GDP was 37 percent larger in 2000 than in 1980.

The rise in primary budget surpluses overcame substantial headwinds. Again, higher real interest rates are an inflationary force, by raising the discount rate for future surpluses or equivalently raising interest costs on the debt. As Figure 2 emphasizes, the US government had to finance debt at large real interest rates for a decade. Taxpayers paid for that. Moreover, investors who bought 10-year bonds at 15 percent yield in 1980, expecting inflation, got repaid in an environment of 3 percent inflation by 1990, a 12 percent windfall real return. Taxpayers paid for that too.

Thus, the end of inflation in the 1980s was a joint monetary, fiscal, and micro-economic reform. In the context of Figure 1, there was a disinflationary shock to the present value of surpluses, a negative of the top panel, largely resulting from greater long-run economic growth. That allowed the additional temporary disinflationary effect of higher interest rates—the bottom panel of Figure 1—to push down inflation quickly and, this time, durably. Without budget surpluses to pay a windfall to bondholders and high interest costs on the debt, the disinflation would likely have failed again.

The reforms of the 1980s came after the monetary tightening, and surely it was unclear that they would be as durable and successful as they turned out to be.

In turn, that uncertainty may help to explain why the recessions of 1980–1982 were so bruising. Interest rates substantially higher than inflation for much of the 1980s may similarly reflect some probability that the policy mix would unravel once again. Latin American history is full of examples in which monetary stringency temporarily stops inflations, but fiscal and microeconomic reforms fail to materialize and inflation spirals upward again (Kehoe and Nicolini 2021). On the other hand, as we shall see, an even better outcome was possible. The ends of hyperinflations and the adoption of inflation-targeting regimes, which installed new fiscal, monetary, and microeconomic regimes more quickly and durably, have quickly stopped inflation and expected inflation without bruising recessions.

The 2000s: Stable Inflation with Large Deficits

Fiscal events turned around in 2000. As seen in Figure 3, debt and deficits grew astronomically. The US public debt/GDP ratio rose from 31 percent in 2001 to 105 percent in 2020. Potential causes include a halving of trend GDP growth in 2000, the recessions of 2000, 2008, and 2020, the allure of low debt service costs, or simple political dysfunction. Long-term projections from the Congressional Budget Office point to steady primary deficits of roughly 5 percent of GDP, followed by worse deficits as aging boomers drain Social Security and Medicare. And those projections assume we don't have another crisis, war, or pandemic.

Yet inflation stayed subdued until 2022. Why? Discount rates are the most natural candidate: Real interest rates, and consequently debt-service costs, went on a steady downward trend, becoming negative for the 2010s. Policy discussion turned to "$r < g$," the possibility that with interest rates permanently below economic growth, government debt never has to be repaid. Olivier Blanchard (2019), in his influential AEA presidential address, argued that "public debt may have no fiscal cost." Calls for cost-free fiscal expansion based on "secular stagnation" and "modern monetary theory" grew. Whether true or not, these views capture a widespread set of expectations that repaying debt would be easy.

At higher frequency, consider the recession of 2008–2009 after the Great Financial Crisis. The deficit widened dramatically, from 1.1 percent of GDP in 2007 to 9.8 percent, 8.6 percent, and 8.3 percent in 2009–2011. Yet inflation declined. Inflation as measured by the Consumer Price Index fell, from a peak of 5.5 percent in July 2008 (at an annualized rate) to 2 percent deflation in July 2009. Inflation stayed below 2 percent for the rest of the decade. Shouldn't those deficits have caused inflation?

Not necessarily. People might have expected that deficits, financing temporary spending to meet an extraordinary crisis, would be repaid by higher surpluses when the crisis was over. The Obama administration promised that debt reduction would follow the stimulus. That's *possible*. However, I don't think that explanation is plausible in this case, and it is not what happened after the fact. Instead, real and nominal interest rates—discount rates, debt service costs—fell like proverbial stones. The federal funds interest rate fell from 5.25 percent in July 2008 to effectively zero, less than 0.25 percent, and stayed there until 2016. The real short-term interest rate

was thus nearly *negative* 2 percent for a decade, and most analysts expected very low interest rates to continue indefinitely. In a quantitative evaluation of events like the 2008 recession (Cochrane 2022a; Cochrane forthcoming, chap. 4), I find that this decline in real interest rates accounts quantitatively for the general pattern that inflation is lower in recessions, despite deficits, and conversely in booms. About two-thirds of all inflation shocks correspond to such discount-rate variation. Similarly, Japan has a debt to GDP ratio of over 200 percent, yet slight deflation. Why? Among other reasons, Japan has had very low real interest rates for three decades.

Why Was There No Deflation in 2008, as There Was in 1933?

In 2008, the world economy experienced the Great Financial Crisis. Many people rightly worried that we would repeat the early 1930s, in which a banking crisis coincided with a disastrous nearly 30 percent cumulative deflation. Why did it not happen again? Many differences between the two episodes have been adduced, but there is an important fiscal difference. Imagine that prices and wages fall 30 percent. Now, government bonds are worth 30 percent more in real terms. The government's nominal tax revenues fall by at least 30 percent, and more as a result of the deep recession. But the government must make the same nominal payments to bondholders. To avoid default, the government must raise taxes or cut spending.

Now, can you imagine our Congress and administration saying in a painful recession that the government must dramatically *raise* taxes, *cut* spending, subject us to "austerity," all to pay an unexpected windfall to bondholders? Isn't the government instead likely to pursue fiscal *stimulus*, to regard deflation as a "bubble," a temporary aberration that fiscal policy should ignore, if only because once the desired reflation occurs, government revenues will recover? That is, of course, exactly what governments did with the minor deflation we observed, as you can see in the large 2008 deficits and larger 2021 deficits in Figure 3. And if that is what people expect, a large deflation cannot happen in the first place. Deflation can only happen if the expected present value of surpluses rises.

The deflation of 1933 included the opposite *fiscal* commitment. The United States was on the gold standard. When the value of gold and currency rose relative to goods and services—deflation—the US government was committed to repay debt in more valuable dollars, which needed tax hikes or spending cuts. By devaluing gold from $20.61 dollars per ounce to $35, and largely abrogating the gold standard in 1933, the US government repudiated this commitment. Deflation stopped immediately (Jacobson, Leeper, and Preston 2019). This new reputation, that the government will *not* raise taxes or cut spending to validate deflation, is exactly, in my reading, why the feared deflation of 1933 did not break out again in 2008.

These episodes clarify a central assumption and theoretical controversy of fiscal theory. Any theory that wishes to determine the price level elsewhere must assume the former sort of "passive" (Leeper 1991) fiscal policy to turn off fiscal theory: Deflation, produced by other means (money growth, or other channels) raises the value of debt. Governments respond with fiscal austerity, which raises the present value of surpluses to match the higher value of debt. Here, I make vivid, and I hope

plausible, the central contrary "active" assumption of fiscal theory: Fiscal policy does *not* respond one-for-one to arbitrary inflation and deflation, paying bondholders whatever the whims of a changed price level require.

The Zero-Bound Era: A Test of Theories

The zero-bound era following the financial crisis of 2008 offers an illuminating test of monetary theories.[6] Interest rates were essentially zero in the United States from 2009 to 2016. The episode was longer in Europe, lasting through early 2020, and longer still in Japan where interest rates effectively hit zero in 1995.

What happens to inflation if interest rates stay at or near zero for many years, and are expected to remain at zero for more years? In these episodes, *nothing*. The pattern of inflation following the 2008 recession was nearly identical to that following the 2000 recession. If anything, inflation at the long zero bound was *less* volatile than in the earlier period.

Existing theories of inflation make clear and contrary predictions. Conventional theories predict a "deflation spiral": Lower aggregate demand produces a recession and lower inflation or even deflation. The Fed lowers interest rates to stimulate the economy. But when interest rates hit zero, the Fed can do no more. Now real interest rates are too high. That further lowers aggregate demand, provoking lower inflation or deflation, which raises real interest rates even more, in a never-ending loop. This longstanding view, based on adaptive expectations, goes back to Milton Friedman's (1968) celebrated AEA Presidential Address, which taught that inflation under an interest rate peg is unstable.

It did not happen. Deflation spirals never broke out. Even in Japan, deflation bumped around 1 or 2 percent at worst.

Rational expectations and New Keynesian models clearly make a different prediction: At the zero bound, inflation becomes *indeterminate* and thus volatile. The interest rate determines expected inflation, but unexpected inflation wanders randomly. In normal times, the Fed can cut off these "multiple equilibria" by threatening to raise or lower interest rates aggressively. But once interest rates hit zero, the Fed is powerless to do so. This is also a longstanding doctrine. Sargent and Wallace (1975) showed that inflation is stable but indeterminate under an interest rate peg with rational expectations. Clarida, Galì, and Gertler (2000) and Benhabib, Schmitt-Grohé, and Uribe (2001) emphasize that this indeterminacy occurs in New Keynesian models when the Fed does not react sufficiently to inflation and at the zero bound.

It did not happen. Again, inflation was if anything less volatile with interest rates stuck at zero than before.

[6]Economists call it the "zero lower bound," because negative interest rates will prompt people to take their money out of banks and hold cash. Several central banks set certain interest rates as low as -1 percent, and the inconvenience of cash kept this "cash arbitrage" from breaking out. Since the bound is not exactly zero, it is sometimes called the "effective lower bound" instead. This section summarizes Cochrane (2018) and Cochrane (forthcoming, chap. 20).

In monetarist thought, the zero bound does not constrain monetary policy. The Federal Reserve can still create reserves or print cash, use it to buy bonds, and let more money (M) in MV = PY raise output (Y) and the price level (P). Starting in 2009, central banks embarked on just such a massive "quantitative easing" program. Bank reserves held at the Fed rose from roughly $10 billion in 2007 to over $2,700 billion by August 2014, an 27,000 percent increase. The monetarist prediction is clear: hyperinflation.

It did not happen. Inflation trundled along a bit less than 2 percent. It is hard to see any effect of quantitative easing in plots of inflation or long-term interest rates. Instead, we learn that money and bonds are perfect substitutes after all when they pay the same rate of interest. Yes, economists continue to debate whether quantitative easing had a few tenths of a percentage point effect on long-term interest rates, but for our purposes the debate is over. A 27,000 percent increase in bank reserves is an atom bomb. If you're debating whether somebody heard a firecracker, it was a dud.

Fiscal theory is consistent with the long quiet zero bound and the silence of quantitative easing. The interest rate target determines expected inflation. Unexpected inflation is determined by news to the present value of future surpluses. Inflation is both stable (no spirals) and determinate (no multiple equilibria or sunspots) at the zero bound. If there is no fiscal or discount-rate *news*, there is no unexpected inflation either. That's not proof: I don't have an independent measure of deficit and discount rate expectations. But fiscal theory is at least *consistent* with the episode, in a way that conventional theories are not. And it's at least plausible that the steady recovery after 2009, combined with very low real interest rates, led people not to worry any more or less than before about debt repayment.

The zero-bound era is thus a powerful experiment. The predictions of classic Keynesianism, New Keynesianism, and monetarism are large and clear, and they fail. Fiscal theory is at least consistent with—not rejected by—the episode. More generally, the zero-bound episode empirically reverses classic doctrines: Inflation can be stable and quiet at the zero bound, and by implication at an interest rate peg greater than zero.

The prediction of long-run neutrality, embodied in the bottom panel of Figure 1, that inflation eventually rises to meet the interest rate, is an inescapable consequence of inflation stability. But "eventually" can take a long time, and it is reasonable that until the zero-bound era central banks and economists never saw interest rates that stayed still long enough, without fiscal news, to observe long-run neutrality. Conversely, the experience of the zero-bound era should make us less nervous about the long-run neutrality proposition embodied in Figure 1.

In the theories, stability comes from rational expectations, which I have used in the model and discussion, more than fiscal theory per se. New Keynesian models are also stable, and inflation eventually rises to meet higher interest rates. One might also pair fiscal theory with adaptive expectations. Such a model will still rule out deflation spirals, but it is likely to produce much different dynamics.

Inflation Rises From the Dead in 2021, and How Will It Be Contained?

Inflation rose suddenly in early 2021, reaching 9 percent by June 2022. From a simple fiscal-theory armchair, this event looks like a helicopter drop, a large version of the fiscal shock plotted in the top panel of Figure 1. From 2020 through 2021, the Fed and Treasury together sent people and businesses checks worth about $5 trillion (Cochrane forthcoming, chap. 21).

But in fiscal theory, a budget deficit is only inflationary if people do not expect it to be repaid by subsequent surpluses. So we must explore why people apparently did not expect the new debt of 2020–2021 to be repaid, in full or in part, and tried to spend it, where they held the new debt of 2008–2020.

There are many suggestive stories. First, politicians and administration officials in 2020–2021 did not emphasize repayment, while deficit reduction promises were a constant theme of the earlier era. Instead, they repeated the view that low interest rates allow for fiscal expansion without worrying about repayment. For example, in her confirmation testimony, just before passage of the $1.9 trillion "American Rescue Plan," Treasury Secretary Janet Yellen (2021) opined that "right now, with interest rates at historic lows, the smartest thing we can do is act big."

Second, I argued above that the sharp decline in real interest rates between 2007 and the decade following 2008 helps to account for the lack of inflation from that era's deficits. But it is unlikely that the 2020s will see an *additional* 3 percentage point or so decline in real interest rates.

Third, the 2020–2021 deficits were much larger than the 2008–2009 era stimulus. Moreover, the pandemic recession was largely a reduction in the economy's productive capacity, not lack of demand. Restaurants were not closed because people didn't have the money to go out to dinner. When supply is constrained, a massive increase in aggregate demand shows up more quickly in prices.

Finally, creating bank reserves and sending checks to people may be more quickly inflationary than borrowing in Treasury markets and spending. Who gets debt matters to how quickly it is spent. Whether the additional debt is in the form of Treasury debt or bank reserves may also matter to expectations of repayment.

How will the current inflation be contained? If, as I suggest, we have suffered a fiscal shock, as shown in the top panel of Figure 1, then in this simple model, monetary policy alone can only reduce that inflation temporarily, by adding the inflation path produced by higher interest rates shown in the bottom panel of Figure 1. Monetary policy can give us lower inflation now, but at the cost of higher inflation later—a form of Sargent and Wallace's (1981) "unpleasant arithmetic." Postponing inflation is still useful: A smaller but longer lasting inflation is desirable in many economic models, as it reduces the disruptive effects of inflation. A Taylor-type rule in which interest rates react to inflation produces such smooth but long-lasting inflation and reduces output volatility.

But in this analysis, monetary policy alone cannot durably eliminate a fiscally-induced inflation. To durably end inflation, monetary, fiscal, and growth-oriented microeconomic policy have to work together, as I argued above they did in the 1980s. And the fiscal headwinds are larger today. In 1980, the debt to

GDP ratio was 25 percent. Now it is 100 percent. A 5 percent real interest rate raises interest costs by 5 percent of GDP, $1.2 trillion, for every year it lasts.

On the other hand, the experience of the zero-bound era and the top panel of Figure 1 say that inflation from a single fiscal shock will eventually die off on its own, even if the Fed does nothing, so long as fiscal policy does not get worse. As inflation did not spiral downward at the zero bound, it does not spiral upward now. This prediction comes also from rational expectations more than fiscal theory per se, but it contravenes conventional wisdom which says the Fed must raise nominal interest rates above inflation to stabilize the latter, and its slowness to act has added greatly to inflation.

This outlook also assumes that we do not have additional adverse fiscal shocks, and that people maintain faith that the US government will pay off the remaining debt. If people did not have faith that the $5 trillion of 2020–2021 debt will all be repaid, will they believe that additional $1 trillion deficits in the next few years will be repaid by later surpluses? If—when—the next crisis hits, and the US government wishes to borrow or print $5 trillion of new debt and money, will people hold the extra debt, or will inflation come even more quickly?

Moreover, debt that is viewed as sustainable because of low interest costs is fundamentally unstable. If investors get scared and demand higher real interest rates, interest costs rise, and the debt becomes unsustainable. Inflation surges, seemingly out of nowhere, or far out of proportion to the initial shock. Abundant fiscal space, a background of healthy long-run surpluses, and financing deficits with long-term debt, which passes higher rates more slowly to interest costs, would squelch these worries. But the US government no longer has that fiscal space.

A Wider and Institutional History

A good theory of inflation should be able to analyze a wide variety of historical episodes and policy regimes, not just postwar US time series. I give a short tour.

The Gold Standard

In an idealized gold standard, the government promises that you can always bring in a dollar and get, say, 1/20 of an ounce of gold. This promise appears to nail down the price level.

The gold standard retains an allure. Monetary policy follows a simple and transparent rule, not requiring divinations by clairvoyant central bankers. The United States didn't even have a central bank in most of the nineteenth century.

But the gold standard is really a case of fiscal theory, not an alternative theory. The government has to have the gold! Governments did not back currency 100 percent with gold, and they certainly never backed debts 100 percent with gold. If they had that much gold, they wouldn't have had to borrow in the first place. If people wanted to turn in a lot of currency for gold, the government had to raise taxes or borrow against credible future taxes in order to get the needed gold.

Currency and nominal debt were backed by the government's ability to tax, not by vaults of gold.

Sims (1999) cites a nice example:

> From 1890 to 1894 in the U.S., gold reserves shrank rapidly. U.S. paper currency supposedly backed by gold was being presented at the Treasury and gold was being requested in return. Grover Cleveland, then the president, repeatedly issued bonds for the purpose of buying gold to replenish reserves. This strategy eventually succeeded.

Cleveland persuaded bond buyers that the United States would run larger future fiscal surpluses, so those buyers were willing to lend.

The final abandonment of gold by the United States in 1971 followed a similar outflow to foreign central banks, who presented US dollars for gold. The Nixon administration was unable or unwilling to take the fiscal steps necessary to buy or borrow gold. In part, it likely did not want to suffer the deflation that restoring gold parity would have implied.

The gold standard is a *fiscal commitment*. The stream of expected future budget surpluses is, on its own, a bit nebulous and hard to forecast, just as stock dividends are hard to forecast. The gold standard offers a commitment of just what that present value will be: The government will raise taxes or cut spending just enough to repay government debt at the gold value (for example, at $20 per ounce), no more and no less. Bond- and money-holders don't need specific surplus forecasts. They just need to understand the commitment and to have a general sense that the government has the fiscal space and the will to do whatever it takes so that the present value of surpluses will, in fact, be the value of government debt.

The gold standard had many flaws, however. First, the gold standard era also featured inflation, devaluation, runs, and crises when governments couldn't or wouldn't tax or borrow to get gold. Those episodes reinforce the fundamentally fiscal nature of the gold standard, and they remind us that all governments have fiscal limits. Second, there was much short-run inflation and deflation under the gold standard. Money does not rise or fall in value relative to gold, but money and gold together can rise and fall in value relative to goods and services. We want a standard that stabilizes the general price level. Third, as above, the gold standard is vulnerable to deflation, when it commits the government to fiscal austerity.

Currency Pegs

A foreign exchange peg is a lot like a gold standard. A government promises to freely exchange its currency, say pesos, for another, say US dollars, at a fixed rate. This peg is likewise a *fiscal* commitment. As with gold, attention often focuses on reserves—how many US dollars the central bank is holding. But as with gold, no country ever has backed all of its government debt with foreign exchange reserves. In the end, the foreign currency peg depends on the government's ability to tax, to get dollars, or to promise future taxes, to borrow US dollars. Even currency boards,

which back *currency* 100 percent, can fail. When the government can no longer borrow to finance deficits, it abrogates the board and grabs its assets. Argentina's (imperfect) currency board, which pegged the peso to the US dollar one for one from 1991 to 2002 with large dollar reserves, fell apart in this way.

An exchange rate peg also suffers the same practical and fiscal problem as the gold standard. A rise in the real exchange rate forces an unwanted deflation and forces the government to validate that deflation with fiscal austerity.

Gold price or foreign exchange *targets* offer some of the fiscal commitment without the run- or speculative-attack-inducing offer to freely buy or sell. But the latter is also a stronger precommitment.

Foreign or Indexed Debt: The Corporate Finance of Government Debt

Many governments issue debt indexed against inflation[7] or borrow in foreign currency. A higher price level does not reduce the real value of indexed debt, nor does a lower exchange rate reduce the real value of foreign currency debt. The government must raise budget surpluses to pay off such debt, or default. Thus foreign or indexed debt act like corporate debt, which must be repaid to avoid default. Domestic currency and nominal (non-indexed) debt act like corporate equity, whose value can fall to meet lower expected profits.

We can then think of the choice between domestic and foreign currency debt, or nominal versus indexed debt, as we think of a corporation's choice between debt and equity. Nominal debt, like corporate equity, allows the government to share the risks of fiscal stress, to let inflation or currency devaluation avoid the pain of formal default. On this basis, for example, Sims (2001) argues that Mexico should not adopt the US dollar. The same argument lies behind the view that countries like Greece should not join the euro.

But equity invites moral hazard. Surpluses are choices, not exogenous shocks. Countries with their own currencies and domestic-currency nominal debt are tempted to inflate rather than to fix fiscal problems. Default is costly, so countries that borrow in foreign currency or indexed debt have an extra incentive to actually run the surpluses in the future that they promised. That pre-commitment allows these countries to borrow in the first place, and at a lower interest rate than otherwise. So, despite the risk-sharing and default-cost reductions of equity, corporate finance recommends widespread use of debt.

Nominal government debt, like corporate equity, works better when government accounts are more trustworthy, and when the country has other means to commit to repay rather than habitually inflate away debt after it has been incurred. The control rights of shareholders are that mechanism for corporate equity. Most naturally, *voters* are that mechanism for government debt. If inflation devalues

[7] An indexed bond protects investors against inflation. In simple terms, rather than get $1 a year from now, an indexed bond pays $1 adjusted upward for any increase in the Consumer Price Index. Naturally, the interest rate for this bond is lower, in a way that adjusts for expected inflation. TIPS, or Treasury inflation-protected securities, are indexed government debt in the United States.

nominal government debt and causes chaos throughout an economy, a lot of voters are really mad.

Thus, the standard ideas of corporate finance suggest that countries with pre-commitment problems, poor fiscal institutions, unstable politics, and untrustworthy government accounts should issue indexed or foreign currency debt or even dollarize. Countries that have alternative pre-commitment mechanisms, strong institutions, and stable democracies with a widespread class of people who prefer less inflation have their own currencies and borrow in those currencies.

Confirming this view, dollarization, currency pegs, and foreign debt are common in the developing and undemocratic world. Successful non-inflating currencies and large domestic-currency debts seem to be the province of stable democracies.

Inflation Targets

In the early 1990s, several countries that were experiencing inflation instituted inflation targets, including New Zealand, Canada, Sweden, and Israel. The targets usually included a formal agreement between government and central bank mandating that the central bank focus on inflation. In these and other cases, inflation dropped on the announcement of the inflation targets and stayed there (for a short history, see Svensson 2010).

Just how were these miracles achieved? Did previous central banks just lack the guts to do what's right in the face of political pressure to inflate? Moreover, just what does the central bank *do* to produce low inflation after the inflation target is announced? One would have thought, and most people did think, that the point is to insulate the bank from political pressure during a long period of high interest rates and painful recession, such as the United States experienced in the early 1980s. But nothing of the sort occurred. Inflation simply fell like a stone on the announcement of the target. Well, "expectations became anchored" when the target was announced, but why? The long history of inflation certainly does not lack for speeches and promises from politicians.

Inflation targets are an agreement between central bank and government. They therefore include an implicit commitment by the government to run *fiscal* affairs so as to pay off debt at (say) 2 percent inflation, no more and no less. Above-target inflation will lead to fiscal tightening. Below-target inflation will lead to stimulus. In this reading, the inflation target is similar to the fiscal commitments of a gold standard or an exchange rate target. But the inflation target aims at inflation directly, not the price of gold or exchange rate, thus eliminating an unwelcome source of relative price variation. The inflation target also avoids the promise to freely trade cash for gold or foreign currency, which can induce runs or speculative attack.

The inflation targets were part of a suite of fiscal, financial, regulatory, and pro-growth reforms. The latter matter: Tax *revenue* equals the tax rate multiplied by income, so the surest way to get more tax revenue is to allow more economic growth. Raising tax *rates* is like walking up a sand dune, since each rise in tax rate lowers income.

The fact that inflation fell quickly after the announcement of inflation targets, without a period of high interest rates or recession, is also revealing. Expected inflation can fall quickly when people see the underlying fiscal problems have been addressed in a durable and credible way.

An inflation target failed instructively in Argentina 2015–2019 (Cachanosky and Mazza 2021; Sturzenegger 2019), one of many cases in which Latin American inflation monetary policies failed because the countries did not solve the underlying fiscal problems (Kehoe and Nicolini 2021). These failures reinforce the point that a successful inflation target is as much a fiscal commitment by the government as a commandment to the central bank.

This implied fiscal commitment is not written in official inflation targeting agreements, nor is it (yet) much discussed in the economic literature. But it surely seems like a reasonable interpretation of the government's side of the deal and the fact of fiscal and microeconomic reforms in successful inflation-targeting episodes. A more formal fiscal rule, announcing how fiscal policy will and won't react to inflation, might anchor expectations more solidly.

The Ends of Inflations

The success of inflation targets echo Sargent's (1982) classic study of the ends of hyperinflations in Austria, Germany, Poland, and Hungary in the early 1920s.

The price level in Germany rose by a factor of 10^{12}. Germany was printing money to finance intractable deficits. Sargent (1982) writes: "Germany owed staggering reparations to the Allied countries . . . considerably larger sums were initially expected of Germany than it ever was eventually able to pay." Germany's hyperinflation stopped instantly when the long-term *fiscal* problem was solved (see Sargent 1982, Figure 2.4). "The Dawes plan assigned Germany a much more manageable schedule of payments." Germany also made permanent reforms to the government budget, in particular firing many unnecessary government and railway workers. With the fiscal problem solved, "Simultaneously and abruptly three things happened: additional government borrowing from the central bank stopped, the government budget swung into balance, and inflation stopped."

The end of Germany's hyperinflation did not involve any monetary tightening. Indeed, Germany printed *more* money. Absent inflation, people are willing to hold a larger real quantity of money. In the similar Austrian case, Sargent (1982) continues, "circulating notes of the Austrian central bank increased by a factor of over 6" after inflation stabilization. There was also no recession: "By all available measures, the stabilization of the German mark was accompanied by increases in output and employment and decreases in unemployment."

In Sargent's (1982) telling, restoring central bank independence was also important, but primarily as a *fiscal* precommitment, to make it harder for the central bank to finance fiscal deficits. After describing how the new Austrian central bank backed note issue by foreign currency and gold, not treasury debt, Sargent continues, that Austrian currency was backed "ultimately by the power of the government to collect taxes . . . by the commitment of the government to run a fiscal policy compatible

with maintaining the convertibility of its liabilities into dollars. Given such a fiscal regime, to a first approximation, the intermediating activities of the central bank did not affect the value of the [Austrian] crown."

Sargent emphasizes that a change in *regime* was necessary for people to believe that the present value of surpluses changed. Announcements, decisions, promises, and temporary and reversible "austerity" policies by today's politicians don't budge long-term expectations. But with a regime change, people's expectations can shift suddenly, not after a slow period of learning by watching inflation itself.

Successful inflation stabilizations involve joint monetary, fiscal, and microeconomic reform, in a durable new regime. They do not have to involve recessions.

Currency Crises

Many currency collapses have clearly fiscal underpinnings. But many large debts and deficits have not led to currency devaluation or inflation. Crises, inflations, and devaluations have also happened in countries with few immediately apparent fiscal problems (Reinhart and Rogoff 2011).

The insight that expected *future* surpluses and deficits drive the value of currency offers a way to understand many episodes. Burnside, Eichenbaum, and Rebelo (2001) study the East Asian financial crises of the late 1990s, in which pegged exchange rates suddenly collapsed. Why? The economies were growing well, the governments did not have substantial debts or deficits, and there was no unusual monetary loosening. However, they show that the crises were precipitated by bad news about *prospective* deficits. Banks had borrowed a lot of short-term foreign-currency debt. The government was poised to bail out banks. A run on banks then became a run on government surpluses.

The episode has broader lessons. The *form* of international investment matters, and endangers the currency when it entangles government finances. Foreign equity investors might decide pull out, and a government may ignore the fact that they sell to locals at low prices. But short-term foreign-currency debt, in banks that the government implicitly or explicitly guarantees, endangers the currency. Contingent, or off-the-books liabilities, such as bailouts of banks, businesses, homeowners, or other debts, make a government more fragile to inflation and devaluation, and are not reflected in conventional surplus and deficit forecasts.

Concluding Comments

The statement that the real value of government debt is equal to the expected present value of surpluses is an ingredient of a theory, not a complete theory by itself. How "fiscal theory" behaves depends on how one fills out the rest of an economic model. In this essay, I place fiscal theory in the context of basic New Keynesian models with rational expectations. One can easily embed fiscal theory in more general models, featuring all the frictions and heterogeneities that make macroeconomics fun and interesting. Better models by which higher interest rates

can temporarily lead to lower inflation are an urgent agenda item. Though such embedding is technically easy, the questions one is led to ask and the results can be quite different, and counter usual intuition.

One might wish for formal tests of fiscal theory. The history of economics warns against that approach. Attempts to construct formal tests of monetarist versus Keynesian versus rational expectations versus real business cycle versus New Keynesian theories *as a class* have never been productive. Instead, each of these models has been evaluated by its ability to plausibly and, later, quantitatively understand episodes and data, and to guide policy, after suitable elaboration. The fiscal theory of the price level will rise or fall in the same way.

Two takeaways: First, monetary and fiscal policy are both important to inflation, as exemplified by the flexible-price case in which the interest rate sets expected inflation and fiscal news determines unexpected inflation, or by the two simulations of Figure 1. Monetary policy is important, as a simplistic reading of "fiscal theory" might not recognize, but fiscal policy also creates inflation that monetary policy cannot fully control, as a simplistic reading of the dictum "inflation is always and everywhere a monetary phenomenon" might deny.

Second, price-level control requires a well-designed monetary-fiscal regime. The present value of the long stream of future budget surpluses *is* somewhat nebulous on its own. As a result, governments create institutions designed to control, communicate, and commit to that present value. The gold standard, currency pegs, independent central banks, inflation or foreign exchange targets, and the hard-won reputations of governments for repaying debts are examples of such institutions. The fiscal theory of the price level leads us to study such institutions or regimes and think about how to improve them.

■ *I thank Erik Hurst, Ed Nelson, Nina Pavcnik, and Timothy Taylor for helpful comments.*

References

Benhabib, Jess, Stephanie Schmitt-Grohé, and Martín Uribe. 2001. "Monetary Policy and Multiple Equilibria." *American Economic Review* 91 (1): 167–86.

Blanchard, Olivier. 2019. "Public Debt and Low Interest Rates." *American Economic Review* 109 (4): 1197–229.

Board of Governors of the Federal Reserve System. 2022. "Federal Funds Effective Rate." Federal Reserve Economic Data. https://fred.stlouisfed.org/series/FEDFUNDS (accessed July 25, 2022).

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo. 2001. "Prospective Deficits and the Asian Currency Crisis." *Journal of Political Economy* 109 (6): 1155–97.

Cachanosky, Nicolás, and Federico Julián Ferrelli Mazza. 2021. "Why Did Inflation Targeting Fail in Argentina?" *Quarterly Review of Economics and Finance* 80 (C): 102–16.

Clarida, Richard, Jordi Galí, and Mark Gertler. 2000. "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory." *Quarterly Journal of Economics* 115 (1): 147–80.

Cochrane, John H. 2018. "Michelson-Morley, Fisher, and Occam: The Radical Implications of Stable Quiet Inflation at the Zero Bound." *NBER Macroeconomics Annual* 32 (1): 113–226.

Cochrane, John H. 2022a. "The Fiscal Roots of Inflation." *Review of Economic Dynamics* 45 (1): 22–40.

Cochrane, John H. 2022b. "Replication data for: Fiscal Histories." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E176381V1.

Cochrane, John H. Forthcoming. *The Fiscal Theory of the Price Level*. Princeton: Princeton University Press.

Friedman, Milton. 1968. "The Role of Monetary Policy." *American Economic Review* 58 (1): 1–17.

Jacobson, Margaret M., Eric M. Leeper, and Bruce Preston. 2019. "Recovery of 1933." NBER Working Paper 25629.

Kehoe, Timothy J., and Juan Pablo Nicolini, eds. 2021. *A Monetary and Fiscal History of Latin America, 1960–2017*. Minneapolis: University of Minnesota Press.

Leeper, Eric M. 1991. "Equilibria Under 'Active' and 'Passive' Monetary and Fiscal Policies." *Journal of Monetary Economics* 27 (1): 129–47.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2011. *This Time Is Different: Eight Centuries of Financial Folly*. Princeton: Princeton University Press.

Sargent, Thomas J. 1982. "The Ends of Four Big Inflations." In *Inflation: Causes and Effects*, edited by Robert E. Hall, 41–97. Chicago: University of Chicago Press.

Sargent, Thomas J., and Neil Wallace. 1975. "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *Journal of Political Economy* 83 (2): 241–54.

Sargent, Thomas J., and Neil Wallace. 1981. "Some Unpleasant Monetarist Arithmetic." *Federal Reserve Bank of Minneapolis Quarterly Review* 5 (3): 1–17.

Sims, Christopher A. 1999. "The Precarious Fiscal Foundations of EMU." *De Economist* 147 (4): 415–36.

Sims, Christopher A. 2001. "Fiscal Consequences for Mexico of Adopting the Dollar." *Journal of Money, Credit and Banking* 33 (2): 597–616.

Sims, Christopher A. 2011. "Stepping on a Rake: The Role of Fiscal Policy in the Inflation of the 1970s." *European Economic Review* 55 (1): 48–56.

Smith, Adam. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Methuen & Co., 1930.

Sturzenegger, Federico. 2019. "Macri's Macro: The Elusive Road to Stability and Growth." *Brookings Papers on Economic Activity* 50 (2): 339–436.

Svensson, Lars E.O. 2011. "Inflation Targeting." In *Handbook of Monetary Economics*, Vol. 3b, edited by Benjamin M. Friedman and Michael Woodford. Amsterdam: Elsevier.

US Bureau of Economic Analysis. 2022a. "Gross Domestic Product: Implicit Price Deflator." Federal Reserve Economic Data. https://fred.stlouisfed.org/series/GDPDEF.

US Bureau of Economic Analysis. 2022b. "Net Lending or Net Borrowing (-), NIPAs: Government." Federal Reserve Economic Data. https://fred.stlouisfed.org/series/AD01RC1Q027SBEA.

US Bureau of Economic Analysis. 2022c. "Real Gross Domestic Product." Federal Reserve Economic Data. https://fred.stlouisfed.org/series/GDPC1.

US Bureau of Labor Statistics. 2022a. "Consumer Price Index for All Urban Consumers: All Items in U.S. City Average." Federal Reserve Economic Data. https://fred.stlouisfed.org/series/CPIAUCSL.

US Bureau of Labor Statistics. 2022b. "Unemployment Rate." Federal Reserve Economic Data. https://fred.stlouisfed.org/series/UNRATE.

Yellen, Janet. 2021. "Opening Statement of Dr. Janet Yellen before the Senate Finance Committee." Speech, United States Senate Committee on Finance, Washington, DC, January 19, 2021. https://www.finance.senate.gov/imo/media/doc/JLY%20opening%20testimony%20%20(1).pdf.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 147–166

Emerging Market Sovereign Debt in the Aftermath of the Pandemic

Kenneth Rogoff

During the COVID-19 pandemic, governments of middle-income emerging market economies faced the same challenges as those of high-income advanced economies: surging health costs, the need to protect populations through lockdowns, and a collapse in global demand. Both sets of countries engaged in many of the same direct and indirect stimulus policies, including cash transfers to individuals, loan guarantees to firms, rent moratoria, and regulatory forbearance for the banking sector. Most central banks in emerging markets cut interest rates, albeit not to zero, and some even engaged in significant quantitative easing.

However, with weaker domestic capital markets, their access to foreign capital notoriously fickle, and generally having to pay much higher interest rates, emerging market stimulus policies stopped well short of the rich-country response. Beyond facing an upward-sloping supply curve for funds, policymakers across the developing world need to be mindful that at any moment, global investor appetite for their country's debt can sharply recede for any number of reasons, including concerns about sustainability (for example Arellano, Bai, and Mihalache 2021; Bianchi, Ottonello, and Presno 2021) and shocks to global interest rates and risk tolerance (Miranda-Agrippino and Rey 2020). Indeed, even with their more cautious policy response, many middle-income countries are now acutely vulnerable to debt crises, high inflation, banking crises, or a mix of all three. During 2020 alone, 44 emerging

■ *Kenneth Rogoff is Professor of Economics and Maurits C. Boas Chair of International Economics at Harvard University, Cambridge, Massachusetts. His email address is krogoff@harvard.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.147.

markets had their debt downgraded. Several smaller emerging markets, including Belize, Belarus, Ecuador, Lebanon, Suriname, Zambia, and Sri Lanka, have already defaulted on their sovereign debt, as has Argentina (again). Among low-income developing economies, which are poorer and typically have much more limited access to foreign private lending, the situation is even worse. According to the World Bank (2022), more than half are either in default or severe debt distress.

The first part of this paper contrasts the fiscal response of advanced economies (in Europe, North America, East Asia, and Oceania) with those of middle-income emerging market countries (a diverse group including countries such as Botswana, Chile, Colombia, Egypt, El Salvador, India, Philippines, Poland, Thailand, Turkey, Tunisia, Ukraine, and Venezuela). The exceptionally low level of global real interest rates, which fell dramatically after the 2008 global financial crisis, has certainly helped most emerging markets immensely. But as we shall see, they still pay significantly higher interest rates than advanced economies, and now face rising interest burdens.

The historical distinction between emerging markets and low-income developing economies was that the former had significant access to global capital markets (punctuated by all-too-frequent debt crises), while the latter had to rely mainly on grants and loans at concessional interest rates. In truth, the line has become much more blurred in the twenty-first century as ultra-low global interest rates have encouraged private lenders to take greater risks. As such, some of the countries classified by the International Monetary Fund as low-income developing economies—for example, Nigeria and Vietnam—share some of the same problems as emerging markets; much of the discussion here applies to them as well. Another confounding factor over the past couple of decades has been an influx of new official lenders, particularly China but others including India and Saudi Arabia, where the creditor may be a government or a state-owned bank, but the terms more akin to a private lender.

The next section introduces the notion of fiscal space, stressing that the difference across economies is more of a continuum than a sharp divide both among countries that have "graduated" from significant risk sovereign default and especially among countries that have not. At one extreme is the United States, which issues the global currency and has enormous borrowing capacity both domestically and internationally, albeit at a cost that depends on long-term global trends for safe real interest rates. At another extreme are poor developing economies who are entirely shut out of global private capital markets. For countries in-between, ranging from Greece to Brazil to Sri Lanka, governments must be sensitive to the concern that investors may lose confidence if they view fiscal policy as unsustainable, resulting in higher, and sometimes abruptly higher, interest rates. A major issue is the extent to which a country must rely on foreign lenders for hard-currency funds, typically denominated in US dollars or euros. Although much of the academic literature employs the convenient assumption that foreign creditors have no legal recourse and the only cost to default is a loss of reputation, I will argue that this approach

misses many of the most fundamental issues in sovereign default, including the jurisdiction of lending (countries have far more control when their debt is governed by domestic courts) as well as the key role of official creditors who typically have very different incentives than private creditors.

Countries whose governments are able to borrow mainly in domestic currency have a distinct advantage over those that borrow in dollars; they always have the option of inflating down their debts. However, the idea that being able to issue debt in one's own currency takes most debt risks off the table is certainly overblown, given that it often takes a very high rate of inflation to significantly reduce the real value of government debt. Among other costs, very high inflation typically wreaks havoc with domestic banks, especially when they are holding—or forced to hold—large quantities of government debt.

We then explore factors that made the decade prior to the pandemic surprisingly benign in terms of emerging market debt distress, even during the 2008–2009 global financial crisis. Important factors include the introduction of independent central banks, the related shift to borrowing in home currency and under domestic law, and the amassing of considerable foreign exchange reserves. The ability of countries to deal with local currency sovereign debt on their own, without involving foreign courts, can make any partial default more efficient, which in turn can help lower the interest rate creditors demand, even adjusting for currency exposure (Du and Schreger 2016). Beyond that, the likelihood of debt crises requiring international intervention is much reduced (Bulow and Rogoff 1988, 1990). However, the ability to borrow significant amounts from abroad under domestic law remains quite limited for both low-income developing economies and lower-middle-income emerging markets.

The last section of the paper takes up the question of whether efforts to assist debt-distressed emerging markets and low-income countries would better be channeled toward outright grants instead of subsidized loans. An important question going forward is whether the time has come to recast the post–Bretton Woods international institutions like the World Bank and to some extent even the International Monetary Fund as aid instead of lending agencies. The need for global cooperation in dealing with environmental and pandemic risks, not to mention the growing size of emerging markets, is likely to undermine the bargaining power of advanced economies. At some point, this evolution of power and incentives will inevitably affect the structure of international financial institutions, and it is none too soon for economists to begin thinking more seriously about this prospect.

Pandemic Response in Low- and Middle-Income Countries

When the global pandemic struck in 2020, policymakers in emerging markets and low-income developing economies faced sharper fiscal constraints than their advanced-economy counterparts. They faced more sustained falls in output and tax revenues, in part due to lack of vaccines, and far more fragile access to foreign capital.

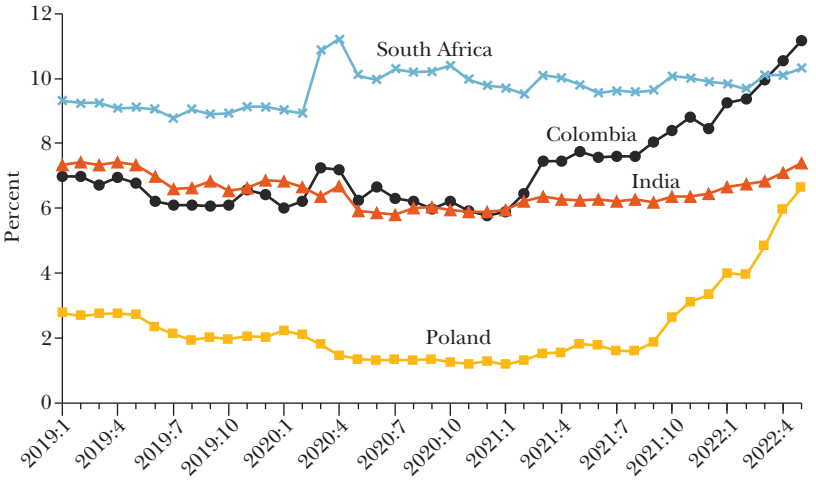
Unprecedented stimulus packages played a large role in cushioning advanced economies, as their public debt/GDP ratios rose sharply from 104 percent of GDP by the end of 2019 to 120 percent of GDP by the end of 2021. In emerging markets, public debt/GDP also rose significantly over the same period (from 55 percent to 66 percent of GDP), but a larger fraction of the rise in the ratio was due to a slower recovery in the output denominator. There is, of course, quite a bit of variance across countries: debt/GDP in Egypt, Sri Lanka, and Brazil at the end of 2021 exceeded 90 percent, while in Angola and India it was over 85 percent. Sri Lanka, of course, defaulted on its debt in 2022.

The ability to successfully engage in any countercyclical fiscal policy at all is a relative novelty for emerging markets. Historically, emerging market fiscal policy has tended to be *procyclical* rather than *countercyclical*, as it is in advanced economies. This is partly because governments in emerging market countries face acute political pressures to spend in good times, but also because their external borrowing constraints are procylical, loosening in booms and tightening in recessions. Particularly in the case of commodity exporters, shocks to export revenues tend to be quite long-lasting, again exacerbating procyclicality (Aguiar and Gopinath 2007). All too often, the combination of these factors has turned what might have been a year-long recession into a multi-year downturn, often including a mix of debt default, financial crisis, and high inflation.

The global financial crisis of 2008–2009 was a distinct exception, with emerging markets recovering relatively quickly. This was partly because the "global" financial crisis was mainly centered in advanced economies, but also due to China's massive construction-centered stimulus that pushed up global demand for commodities. The pandemic, however, has been more challenging.

In contrast with advanced economies, which have enjoyed a falling interest burden despite higher debt levels, many emerging markets and low-income developing economies seen their interest burden rising (Kose et al. 2022). This is in part because their interest rates did not fall by as large a fraction, in part because their outputs have recovered more slowly, and in part because of a shorter maturity structure; a country such as the United States is still benefiting from the fact that it can roll over, say, 30-year Treasury debt, at much lower interest rates than three decades ago. As of mid-2022, interest rates on long-term domestic sovereign debt exceeded 10 percent in over one-third of emerging markets, compared to around 3 percent in the United States, and 1 percent in Germany. Figure 1 shows representative long-term rates on domestic debt for Colombia, India, Poland, and South Africa from January 2019 to May 2022. Similarly, the risk premium on external foreign-currency debt (excess interest rate over safe debt) exceeded 10 percent for one-quarter of emerging markets (IMF 2022a). The latter is hardly surprising, given that external debt service alone absorbs more than one-third of export revenue for many countries. Figure 2 gives private plus public external debt interest payments relative to exports for a grouping of large emerging markets as well as for a group of smaller emerging markets. As the figure illustrates, these countries had entered the 2008 global financial crisis in a relatively

Figure 1
Long-Term Interest Rates for Each Country: January 2019 to May 2022



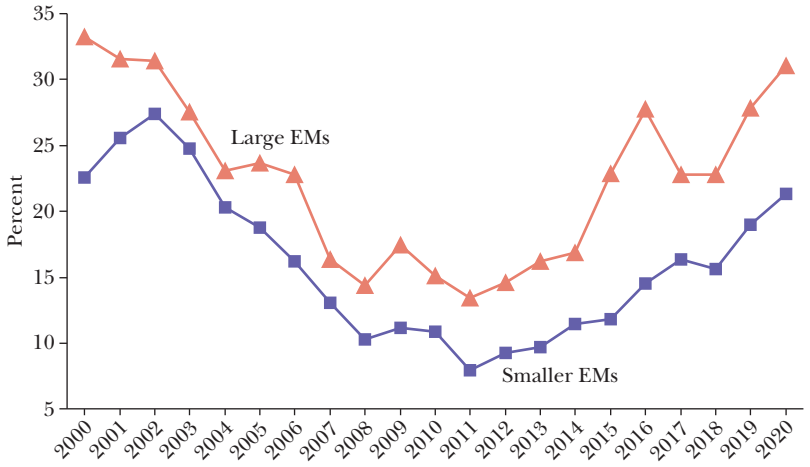
Source: Rogoff (2022).
Note: Data are at monthly frequency. Long-term interest rates are measured by ten-year government bond yields for all four countries.

strong position, having deleveraged after their earlier debt crises, but in the run-up to the pandemic and continuing into it their external debt positions had become more vulnerable.

Fiscal Space and Fiscal Policy

Emerging market countries that rely significantly on foreign capital for government borrowing and private investment must always be mindful of the risks of "sudden stops," where international credit freezes up or becomes available only at prohibitively high interest rates. This is particularly the case when the maturity structure of debt is relatively short-term, so that the government is constantly pressed to roll over a significant share of its debt every year. It can also be a problem when a country is running a large current account deficit, implying that it is quite dependent on a steady flow of fresh foreign financing. Historically, there have been many waves of sovereign debt and financial crises across the world, with most of today's advanced economies having defaulted several times across the centuries. While such economies have been able to avoid defaulting since World War II, with only a couple of exceptions (Reinhart and Rogoff 2009), quite a few have had to resort to International Monetary Fund programs. For example, the United Kingdom famously kept calling on the IMF for help until the 1980s. It should be emphasized that default on

Figure 2
Average Ratio of Total External Debt Service over Exports of Goods, Services, and Primary Income, for Each Group of Countries: 2000–2020



Source: Rogoff (2022).
Note: "EMs" refers to emerging markets. The group "Large EMs" consists of Brazil, Indonesia, Mexico, Nigeria, South Africa, and Turkey. The group "Smaller EMs" is defined as Algeria, Colombia, Egypt, Kenya, Pakistan, Peru, Philippines, and Thailand. The average ratios (in percent) for "Smaller EMs" from 2000 to 2004 do not include Algeria in the calculation since Algeria's data are missing for those years. Otherwise, the average ratio is an unweighted average of total debt service percentage across all countries in the group.

sovereign debt tends to be partial, not total (Bulow and Rogoff 1988), as countries do not go out of business; particularly if debt is adjudicated in foreign courts, it does not completely disappear simply because a debtor country says so.[1]

By contrast, emerging markets have experienced waves of debt distress and/or default in the 1980s (centered in Latin America but extending worldwide) and in the late 1990s (centered on Asia), with many other prominent cases around these (for example, Mexico in 1994, Russia in 1998, and Argentina in 2002). Most emerging markets have experienced multiple episodes where they have either defaulted outright on their foreign debts or, in modern times, were forced to get emergency lending from the International Monetary Fund.

The extent to which swings in market confidence are mainly the result of unsustainable fiscal policy versus a self-fulfilling loss of confidence has been widely

[1] The canonical model of Eaton and Gersovitz (1981), which is widely used in the theoretical literature, implicitly assumes that creditors have all the bargaining power and can impose on a country repayment equal to the maximum penalty that can be extracted, regardless of cost to the creditors. This model offers a useful starting point, but clearly not a very realistic description of real-world debt negotiations. Yue (2010) attempts to combine the two classes of sovereign debt models.

debated in the literature for decades (Obstfeld and Rogoff 1996). Loss of investor confidence can be self-fulfilling if investors begin demanding such high interest rates that debt becomes unsustainable, even if it might have been sustainable at lower interest rates. However, whereas the theory may be ambiguous, the empirical evidence is abundantly clear: countries with higher debt/income levels (including hidden "off-the-books" debt) and a history of serial default are distinctly more vulnerable to debt crises (Reinhart, Rogoff, and Savastano 2003). For vulnerable countries, a significant rise in global real interest rates or a fall in the prices for their main commodity exports are the most common triggers, but other factors such as domestic political instability can play a role. The Latin American debt crisis of the 1980s (which extended to other parts of the world as well) is a canonical example where high global real interest rates, low commodity prices, political instability, and loss of investor confidence all combined to create a perfect storm.

Countries that have limited fiscal space by virtue of their history, politics, and institutions need to balance the costs and benefits of fiscal stimulus very differently from countries that issue debt regarded by investors as "safe," even in a global crisis or recession. The classic model of Barro (1979), which has governments issuing debt to cover large, unexpected costs, does not straightforwardly apply if high debt results in loss of market confidence. Indeed, the fiscal multiplier can become negative at very high debt levels with the impact of higher interest rates trumping any direct benefit from stimulus.

Thus, when emerging economies are faced with a crisis such as the COVID-19 pandemic, they face an uncomfortable trade-off between the short-term benefits of fiscal stimulus and the cost of being seen as a higher-risk borrower. Importantly, even if interest rates on the country's debt do not rise immediately after a large stimulus, most models suggest the odds of a sudden stop happening eventually will go up, and interest rates on their debt rises accordingly (Fang, Hordy, and Lewis 2022). Bianchi, Ottonello, and Presno (2021) develop a quantitative framework for analyzing the trade-offs between short-term stimulus and medium-term debt rollover risks; they show that for plausible parameters, fiscal restraint may be called for even when fiscal multipliers appear to be quite high. Arellano, Bai, and Mihalache (2021) calibrate a quantitative model for emerging markets for the COVID-19 epidemic and show that the constraints posed by lack of fiscal space is acute. Several earlier studies, including Ilzetzki, Mendoza, and Végh (2013) and Kose et al. (2021), have shown that medium-term fiscal multipliers are negative for countries that enter a recession with very high debt levels (see also Ostry, Ghosh, and Espinosa 2015).

It is worth noting that although the focus of this article is on emerging markets, who as a group face significantly greater debt risks than advanced economies as a group, investors do not regard all advanced economies as issuing equally "safe" debt. The United States, as previously noted, is in a league of its own. Approximately two-thirds of global central bank reserves are held in US dollars and a large share of global imports and exports are priced in dollars. When private financial and nonfinancial firms borrow abroad, an outsized share is priced in dollars. By some measures, the US dollar is the anchor or reference currency for a larger share of

the global economy today than it was under the fixed rate Bretton Woods system of the 1950s and 1960s (Ilzetzki, Reinhart, and Rogoff 2020, 2022). Policies of the US Federal Reserve have an enormous influence on the global financial cycle. It is true that US bond rates are quite similar to those of major eurozone countries after adjusting for exchange rate expectations (Du and Schreger 2022). However, the US economy dominates global public and corporate debt markets, having placed in international markets roughly as much of both types of debt as the other major advanced economies combined.

Moreover, whereas virtually all the advanced-economy governments are able to borrow in their own currency, the United States is unique in that the overwhelming share of US corporate debt held abroad is also in US dollars. For other countries, a significant share of corporate bonds held by foreigners is in another currency (typically US dollars). For smaller advanced countries such as Australia, New Zealand, and Canada, the share of corporate debt held abroad denominated in foreign currency exceeds 90 percent; even for the United Kingdom, 75 percent of corporate bonds held by foreigners is in foreign currency (Maggiori, Neiman, and Schreger 2020).

The euro area is second after the US dollar in its global debt footprint, but much smaller by most measures. Although some may argue that countries such as Greece, Portugal, and Italy now issue completely safe debt, one might also argue that investors consider their debt completely safe only because Northern eurozone countries implicitly stand behind it. These three countries all experienced severe debt crises less than a decade ago, with Greece having defaulted on its government debt as recently as 2012. During the pandemic, Italy's rates again began to spike until Europe unveiled its €700 billion "next-generation EU" which involved massive transfers from Northern to Southern Europe in the form of outright grants (about half) and low-interest-rate loans. Certainly, overall euro area fiscal space appears ample in a very low interest rate environment, but it remains to be seen whether Europe's political fabric will fray if real interest rates revert toward the very long-run trend.

This digression is meant to underscore that the US exceptionalism in debt capacity does not necessarily extend equally to all other advanced economies. Although most advanced economies have "graduated" from outright default on government debt, this does not necessarily imply full immunity from market pressures that discourage countries from engaging in countercyclical fiscal policy as forcefully as they might otherwise. The fact that countries with very high debt levels (for example Italy and Greece) tend to have lower growth levels (Reinhart, Reinhart, and Rogoff 2012) is not likely a random coincidence, but rather a result of having fiscal policy partly shackled by concerns of risking, at some point, a painful interest rate rise.

Finally, the fact that interest rates on government debt are often lower than growth rates by no means implies bountiful fiscal space for emerging markets. As Mauro and Zhou (2021) show, between 1865 and 2018, the mean percent of years that a negative interest rate growth differential has prevailed in emerging market countries was 75 percent (in part due to financial repression which helped to hold

down interest rates in those countries). The gap between the government borrowing rate and the economic growth rate gives a country room to run a (small) deficit, but if primary deficits are large enough, debt/GDP ratios can balloon to unsustainable levels anyway. In emerging markets, both the interest rates they face and the growth rates they experience are extremely volatile, so although the interest rate growth differential may typically be negative, it can turn large and positive at short notice. Indeed, historically, negative interest rates growth differentials frequently prevailed just prior to a crisis. Interest rates and growth rates are also volatile in advanced countries, but the variance of each is significantly lower.

Risks of a Systemic Emerging Market Financial Crisis: Two Ways to Tell the Story

The COVID-19 pandemic has shaken what had been a remarkably quiescent period in emerging markets debt. The fact there had not been a systemic outbreak of emerging market debt crises for nearly two decades[2] had been somewhat surprising, given that commodity prices remained extremely volatile. For example, the price of a barrel of oil, arguably the world's most important traded good, collapsed from $114 in June 2014 to under $30 in January 2016 without triggering a default in major emerging market oil exporters like Russia, Mexico, or Nigeria. Similar commodity price volatility has been repeated over the course of the pandemic. True, there had been debt restructurings in Argentina (2019) and Ukraine (2015) during this benign period, but otherwise only a handful of smaller generally low-income countries were forced to renegotiate external debts, including Mozambique, Belize, Mongolia, Chad, Grenada, and Ecuador. Turkey, Brazil, and Indonesia, despite periods of extreme duress, did not.

Unfortunately, two relatively calm decades is hardly enough to demonstrate that emerging market economies have "graduated" from sovereign default to the same extent as advanced economies; there have been relative calm periods in emerging market debt before. For example, after the great wave of emerging market debt crises in the Great Depression of the 1930s, there were only scattered defaults among emerging market for several decades: for example, Argentina defaulted in the 1950s, Brazil, Chile, Peru, Costa Rica, and Zimbabwe all defaulted in the 1960s, and Chile again in the 1970s. However, after this relatively calm period, major waves of defaults broke out in the 1980s, 1990s, and early 2000s.

Regardless of how the 2020s unfold, the extended period before the pandemic was sufficiently unusual that it is worth asking whether it owed mainly to fundamental changes in emerging market debt management, or if the quiescence was mainly due ultra-low global interest rates.

[2]The European debt crisis of 2011–2012 hit periphery mid-tier advanced economies such as Spain, Ireland, and Italy plus the lower-tier advanced economies of Greece, Portugal, and Cyprus.

Perhaps the most important cause for optimism has been the strengthening of central bank independence across the world.[3] Four decades ago, only a small handful of central banks in emerging markets had substantial independence. Today, or at least until the pandemic, some degree of operational independence has been the norm, with the approach proving remarkably successful even in emerging markets—albeit not without backsliding in some cases. (The serial firing of central bank governors by Turkish President Recep Tayyip Erdogan was surely a major contributor to that country's inflation rate of over 70 percent in 2022.)

Mainly as a result of greater central bank independence and the resulting fall in trend inflation, many emerging markets have been able to deepen their domestic capital markets dramatically. In turn, this has allowed governments of these countries to borrow greater amounts in their own domestic currency, even from foreign lenders. During the 1980s and 1990s, inflation at annual rates measured in thousands of percent, or more, had amounted to a de facto default on domestic currency bonds, and it was only when inflation stabilized that it became possible to substantially rebuild domestic capital markets. But in Brazil, for example, almost 70 percent of government debt held by foreigners was denominated in Brazilian real; for Mexico, a similar share is held in Mexican pesos (Du and Schreger forthcoming).

The shift from foreign to domestic courts as the venue for decision-making about debt relief issues is also quite significant, regardless of currency of denomination. When debts are litigated in debtor country courts, foreign creditors have significantly less leverage to engage in "holdup" actions that cause economic damage not only to the debtor, but also to trading partners outside the country. This is sometimes referred to as "the free rider problem." This reality is consistent with bargaining-theoretic models of sovereign debt that are rooted in each party's legal rights; in models where the incentive to repay is based mainly on reputation for repayment,[4] court jurisdiction presumably would matter little. Following the logic of bargaining-theoretic models, Bulow and Rogoff (1990) argue that steps taken by debtor countries to strengthen their own domestic debt market institutions could help resolve the inefficiencies involved in debt crises, including the long multiyear delays that are common in resolving them, and thus could mitigate the effects of such crises. (In some cases, even where jurisdiction is in domestic courts, foreign governments may have significant leverage, as was the case in the 1994 Mexican debt crisis where ink was not yet dry on the North American Free Trade Agreement.)

There has also been a massive increase in central bank reserves, especially in Asia but also in many African and Latin American countries. China famously has over $3 trillion dollars in reserves, but among prominent emerging market examples, India and Russia have over $600 billion, and Brazil $300 billion (IMF 2022b). Reserves cannot indefinitely plug up a gaping fiscal hole, but they do serve as a

[3] Economic theory began to emphasize the importance of central bank independence in the mid-1980s. As one example, Rogoff (1985) presents a model central bank independence as a useful device for achieving anti-inflation credibility, including via inflation targeting; see also Alesina and Summers (1993).
[4] Eaton and Gersovitz (1981) is the classic reference.

backstop for dealing with liquidity crises that may arise when there is significant external borrowing (including private sector borrowing).

The shift away from rigid exchange rate pegs has also helped. During the late 1990s and into the early 2000s, there were many cases where a rigidly fixed exchange rate (often to the US dollar but in the case of 1990s Europe to the German deutschmark) created a lightning rod for speculative attacks (Obstfeld and Rogoff 1995). Typically, in such an attack, a large number of speculators short sell the currency, up to a point where the central bank can no longer support the fixed value. The speculators can cash in when the fixed rate fails and the currency depreciates. When a country's central bank tries to hold on too long to an exchange rate that is unsustainably high, it can lose billions of dollars in the blink of an eye. Today, however, many emerging market economies have shifted to softer exchange rate pegs, which perhaps introduces sufficient uncertainty into the policy reaction function that it becomes more difficult for speculators to coordinate on an attack. (This is only a conjecture; the question merits further research.)

Especially since the global financial crisis, there is now vastly greater transparency about public borrowing, which reduces the odds a country will be able to borrow excessively. For example, the event that catalyzed the European debt crisis in 2010 was the revelation that the Greek government had borrowed far more than its official numbers showed, which is not historically an uncommon problem in debt crises. Until the 2008 global financial crisis, the IMF and the World Bank did not maintain any kind of long-dated domestic debt database that might be used for academic studies of debt vulnerabilities. Since then, building on the extensive historical public debt data bases first developed by Reinhart and Rogoff (2009), the International Monetary Fund and the World Bank now provide much more complete data on public (and some extent private) debt; Abbas and Rogoff (2019) describe the modern literature on debt databases.

Lastly, and perhaps somewhat speculatively, the expanded availability of central bank swap lines, mainly due to the US Federal Reserve, may be helping mitigate crunches in US dollar liquidity. True, only a few emerging markets were given swap lines during the pandemic (because the Federal Reserve must be mindful of not taking on too much default risk); nevertheless, flooding the global market with dollar liquidity may have helped markets more broadly.

Unfortunately, against this substantial list of factors that offer hope that some emerging markets might have "graduated" from sovereign default fears, there is a perhaps even more impressive list of reasons why risks still abound in the post–COVID-19 era. Some of these factors have already been discussed, but are worth re-emphasizing here.

First, many emerging market countries have record levels of government debt relative to income, and in contrast to advanced economies, the interest burden of debt for these countries has been rising, with the ratio of debt service to exports (including on private borrowing) reaching new peaks as well.

Second, external debt of emerging market countries (again including public and private) is extremely elevated, having already reached record levels in 2019

prior to the pandemic (Kose et al. 2021). Although private loans are not explicitly government guaranteed, widespread private-sector default on foreign currency debt often creates irresistible pressure for government bailouts—and thus this foreign currency debt can reasonably be viewed as a form of "hidden debt" that often sharply impacts government balance sheets. Emerging market governments can no more ignore debt distress in large systemically important corporations than can advanced economy governments.

Third, although countries have been more successful in issuing domestic debt, during periods of stress the government almost invariably forces the domestic banking sector to absorb the debt; this was a major feature of the European debt crisis that began in 2010 and took years to stabilize. This creates the infamous "doom loop": as bank profits collapse, the government is forced to bail them out, which in turn makes the government debt vulnerable, which in turn weakens banks further, and so on. Historically, financial crises in emerging markets have often been followed by sovereign debt crises (Reinhart and Rogoff 2011).

Fourth, although emerging market governments are now more often able to borrow a much larger fraction of debt in domestic currency, this does not shield countries from the risk of a destabilizing loss of confidence that can lead inflation and domestic currency interest rates to spike. Central bank independence is to some extent a bulwark, but the risk of a scenario in which a central bank is pressured or forced to accommodate fiscal dominance remains significant. In much of the world, there is tremendous political pressure to deal with inequality. There may be good arguments for this shift, but anti-austerity policies in financially fragile emerging market economies at a time of weak growth carry significant risks.

Dealing with a Debt Crisis

What are the options if an emerging market country finds itself in a situation where it is running into debt problems, with soaring market interest rates and massive capital outflows?[5] A heterodox mixture of policies, including default, inflation, and financial repression, have long been commonplace in emerging markets as well as in advanced economies (Qian, Reinhart, and Rogoff 2011). Financial repression refers to policies that make it easier for the government to borrow by restricting the ability of domestic investors to access alternatives, including requirements that financial players like banks, insurance companies, and pension funds hold reserves in domestic currency, or policies that raise the costs of investing in foreign assets. Financial repression is a form of implicit taxation, falling mainly on captive domestic residents.

Orthodox solutions to a sovereign debt problems typically involve cutting budget deficits through higher taxes or lower spending, which have come to be

[5] Kose et al. (2022) look at options for reducing unsustainable debt. For an extensive assessment of the current literature on sovereign default, see Abbas, Pienkowski, and Rogoff (2019).

described as "austerity policies," although in the canonical "sudden stop" scenario some belt-tightening is inevitable if foreign creditors are unwilling to continue lending and official bailouts do not fully fill the gap. Default is a very real option that governments have turned to countless times over the centuries, although contrary to popular rhetoric, this will not eliminate the need for austerity in cases where the country is running large government and trade deficits and external private funding has dried up (such as Greece in 2010).

Where do official lenders fit in, and in particular the international agency charged with maintaining global financial stability, the International Monetary Fund (IMF)? Traditionally, IMF policies have been aimed at making orthodox adjustment policies less painful. They do so in two ways. First, the IMF provides short-term loans to help partially fill the financing gap that arises when a country is cut off from market lending or forced to pay prohibitively expensive interest rates. The IMF's short-term loans are typically at very favorable interest rates, comparable to the rates faced by "safe" advanced-economy borrowers; moreover, the IMF often coordinates supplemental loans from other official creditors such as the World Bank or bilateral arrangements with other governments. IMF loans rarely fully offset the budget shortfall caused by the collapse of private credit, but they help to mitigate the short-term pain. Second, and equally importantly, the IMF tries to help the debtor country enunciate an orthodox policy plan that private markets find credible, thereby helping to restore the flow of private funds, which for most middle-income emerging markets are far larger than official funds.

One long-standing critique of International Monetary Fund programs is that, with its emphasis on facilitating orthodox adjustment and consequent avoidance of default of any type, the programs are overly focused on ensuring that foreign private creditors get paid in full and on time. Yet in some cases, particularly where the inherited debt burden is exceptionally large, it is by no means obvious that the debtor country's low- and middle-income citizens fare better under an IMF-style adjustment plan than they would do in an outright default, or equivalently a rescheduling of repayments that lowers the market value of the debt. Although academics have long stressed this point, it has only recently started to gain traction in policy circles. This is perhaps because multilateral loans, and IMF loans in particular, can only be made at the request of the incumbent debtor country government, which is often keen to avoid a politically destabilizing default. It is also important to remember that private creditors have significant leverage with the multilateral lending institutions through their shareholder governments, where the major creditor countries have a large voice.

The International Monetary Fund has long been cognizant of this concern, most famously expressed in First Deputy Managing Director Anne Krueger's (2002) proposal for a mechanism to facilitate sovereign bankruptcy. The bar for implementing Krueger's proposal, however, was high—indeed, it required a new international treaty, which did not come to pass. In recent years, especially since the IMF's much-criticized bailout of Greece during the European debt crisis, the IMF has tried an informal approach to achieve the same end. In principle, the IMF now

claims that it will no longer make loans into situations where overall debt levels are clearly unsustainable and some kind of debt write-down is otherwise inevitable. The idea is that by making this pre-commitment, the IMF hopes to prod private creditors to offer a significant debt write-down in order to catalyze bailout funds from which they, too, will benefit.

This International Monetary Fund policy shift is potentially an important change, although in stating that it will be "tougher" with private creditors it cannot shake off the fact that it cares about the welfare of the debtor country. The IMF's empathy is laudable, but can be a distinct disadvantage in bargaining (Bulow and Rogoff 1988). In practice, the problem is that assessing sustainability remains far from an exact science, which in turn creates pressures on the IMF to be overly optimistic. Indeed, Schlegl, Trebesch, and Wright (2019) show that if one looks in detail at bond payment receipts over the past five decades, the IMF may indeed be clearly senior to other *official* creditors, but de facto, the same cannot be said for the official sector as a whole relative to private creditors. Bulow, Rogoff, and Bevilaqua (1992) reach a similar conclusion using both a theoretical bargaining model and cross-country regressions on sovereign debt secondary market prices for 1980s debt crisis countries, where a higher official debt share does not seem to be associated with lower secondary market prices in deeply debt distressed countries. They give examples where, in practice, other official lenders make loans to pay off the IMF, and then aid funds are diverted to pay off the other official creditors.[6]

In light of the ability of debtors and creditors to game the multilateral financial institutions, does the financial structure of institutions like the International Monetary Fund, heavily geared toward lending rather than grants, still make sense? One problem with having most aid be in the form of loans is that largest packages tend to go to wealthier middle-income countries—for example, the large IMF packages given to Greece during the European debt crisis and to Argentina in 2018. If funds now dedicated to debt relief instead came in the form of outright grants, with no repayment required, perhaps a higher proportion might go to the poorest countries, as Bulow and Rogoff (2005) argue. Recently, Okonjo, Tharman, and Summers (2021) have forcefully revisited the grants versus loans debate and conclude that climate change and pandemics represent global public goods that are best dealt with through grants, not loans.[7] Obviously, a drawback to funding through outright transfers is that the resources need to be replenished constantly, and advanced economies must be willing to do so. This approach would be in contrast to the current structure of the IMF, which in principle is a rotating fund, where repayments of old loans replenish the ability of the IMF to make new loans in the future. In the past,

[6]If the official sector were senior, then the price of private debt should be negatively correlated with the share of official lenders in the country's debt. In fact, controlling for global interest rates and other factors, and using an instrumental variables approach, Bulow, Rogoff, and Bevilaqua (1992) find that the correlation is positive.

[7]Rogoff (2019) argues for the creation of a World Carbon Bank, which would channel aid and technology from rich countries to lower-income countries, for example, to help facilitate reduced emissions from coal plants—or even the outright removal of such plants.

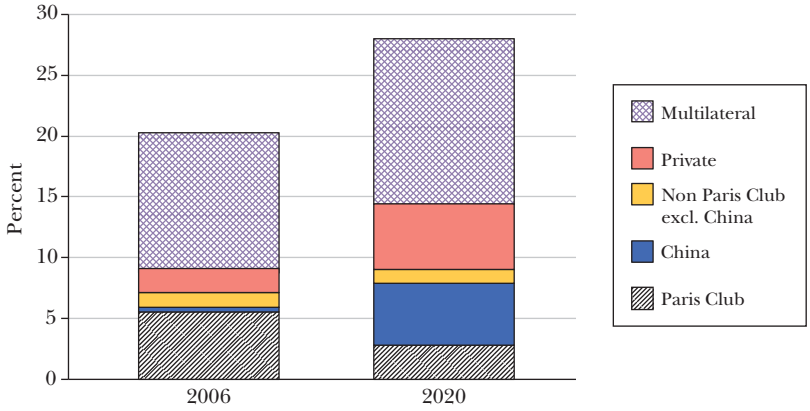
having high-income countries give substantial amounts of aid on a sustained basis has not been a viable political equilibrium. In the coming decades, however, this could change. Given concerns over the environment, international migration, and global health, the bargaining position of the developing world is arguably strengthening, potentially making larger quantities of aid more sustainable.

It is true that the International Monetary Fund already has an aid-like instrument at its disposal called "special drawing rights" (SDRs), which can be issued with a sufficient majority vote. The IMF issued $650 billion worth of special drawing rights in 2021 in response to the pandemic, and close to $300 billion in 2009 to deal with the global financial crisis. Space limitations prevent giving the topic full justice here, but suffice to say that SDRs were designed as an international liquidity instrument and not as an aid vehicle. First, by treaty, the funds go to countries in proportion to their quotas in the IMF, which has the effect of giving only a small fraction to the poorest countries. Wealthier countries can transfer their individual quotas to poor countries, but of course this can be done without SDRs in a way that would be more transparent to donor country taxpayers, and hence more sustainable. Second, SDR funding is government-to-government with essentially no conditionality. Because many poor countries have weak institutions, it is difficult to be sure it works to alleviate poverty as opposed to say, financing capital flight, corruption, and loan repayments to private creditors. Using the SDR as an emergency aid vehicle in a world crisis when there is no other option can make sense, but a durable aid mechanism requires some better way of allocating across beneficiaries and of imposing conditionality.

As noted at the outset, the lines dividing economies from advanced to emerging to developing should be thought of as a continuum, and many low-income developing economies today are experiencing symptoms of the debt crises akin to what emerging markets have long faced. Both private creditors and new official creditors (mainly but not exclusively China) have come to play a larger role; the problems in resolving these unsustainable debts and providing fresh aid have become increasingly difficult. Figure 3 shows a breakdown of the loans outstanding to developing economies as of April 2022. Note how China has increasingly supplanted other bilateral official lenders, while at the same time the multilateral lenders have considerably expanded their footprint.

In the case of these relatively poor countries, the overall size of the loans from multilateral lenders remains small. During the pandemic, the G-20 intergovernmental forum was able to reach a "debt servicing suspension agreement" to provide temporary relief from official loans, but it has been difficult to force private creditors to follow or to agree on actual write-downs of the debt, which, in all likelihood, will ultimately be necessary. Naturally, creditors who are willing in principle to partly forgive loans do not want the relief they provide to end up helping pay off other, tougher-minded creditors. This is a version of the "free rider" problem discussed earlier, but this time also involving official lenders. The problems facing the world in giving pandemic debt relief to very needy poorer countries likely portends considerable problems for any larger emerging markets that run into debt problems, especially given the much greater sums at stake.

Figure 3
External Debt Obligations to GDP for DSSI-Eligible Countries, by Groups of Creditors: 2006 versus 2020



Source: Chabert, Cerisola, and Hakura (2022).
Note: DSSI stands for the Debt Service Suspension Initiative of the World Bank. DSSI-eligible countries are a group of 73 low-income developing economies that is eligible for suspension of servicing on official debts during the pandemic. PPG stands for "public and publicly guaranteed" debt, whether it is issued by a public or private lender.

Conclusion

Since the global financial crisis of 2008–2009, the bulk of academic research on debt and fiscal policy has focused on countries where government debt is assumed to be extremely safe and repaid with extremely high certainty. But this research primarily applies to high-income advanced economies, and perhaps only to the largest and richest ones at that. For emerging markets and developing economies, fiscal space is a very real constraint that can surface under a variety of circumstances, including rising world interest rates, falling commodity prices, or a global recession. With emerging markets and developing economies straining under the burden of the pandemic, in addition to price spikes for imported food and energy, the risks of systemic middle-income country debt crises are real and place a significant constraint on policymakers. For poorer developing countries, the debt problem has already arrived. Finding ways to mitigate that constraint is a major challenge for debtor countries and multilateral lenders.

Near term, making sure that troubled debtor countries are aware of the full menu of options, including both orthodox and heterodox strategies, is important. Finding faster ways to achieve debt write-downs for countries facing unsustainable debt burdens is a perennial problem that has again become front and center for

world policymakers. Longer term, a rethinking of the Bretton Woods financial institutions to incorporate a greater emphasis on outright grants instead of loans makes more sense than ever, given the growing importance of global public goods and the growing importance of the developing world in addressing today's most pressing global problems.

■ *The author is grateful to Jieying Zhang for excellent research assistance, and to the Molly and Dominic Ferrante Fund at Harvard University for research support. Erik Hurst, Nina Pavcnik and Timothy Taylor provided extremely helpful comments on an earlier draft.*

References

Abbas, S. Ali, Alex Pienkowski, and Kenneth S. Rogoff, eds. 2019. *Sovereign Debt: A Guide for Economists and Practitioners.* Oxford: Oxford University Press.

Aguiar, Mark, and Gita Gopinath. 2007. "Emerging Market Business Cycles: The Cycle is the Trend." *Journal of Political Economy* 115 (1): 69–102.

Alesina, Alberto, and Lawrence Summers. 1993. "Central Bank Independence and Macroeconomic Performance: Some Comparative Evidence. Journal of Money, Credit and Banking 25 (2): 151–62.

Arellano, Cristina, Yan Bai, and Gabriel P. Mihalache. 2021. "Deadly Debt Crises: COVID-19 in Emerging Markets." NBER Working Paper 27275.

Barro, Robert J. 1979. "On the Optimal Determination of Public Debt." *Journal of Political Economy* 87 (5): 940–71.

Bianchi, Javier, Pablo Ottonello, and Igancio Presno. 2021. "Fiscal Stimulus under Sovereign Risk." Unpublished.

Bulow, Jeremy, and Kenneth S. Rogoff. 1988. "Multilateral Negotiations for Rescheduling Developing Country Debt: A Bargaining-Theoretic Framework." *IMF Staff Papers* 35 (4): 644–57.

Bulow, Jeremy, and Kenneth S. Rogoff. 1990. "Cleaning Up Third-World Debt Without Getting Taken to the Cleaners." *Journal of Economic Perspectives* 4 (1): 31–42.

Bulow, Jeremy, Kenneth S. Rogoff and Afonso Bevilaqua. 1992. "Official Creditor Seniority and Burden Sharing in the Former Soviet Bloc," *Brookings Papers on Macroeconomic Activity* (1): 195–222.

Bulow, Jeremy, and Kenneth S. Rogoff. 2005. "Grants versus Loans for Development Banks." *American Economic Review* 95 (2): 393–97.

Bulow, Jeremy, Carmen M. Reinhart, Kenneth S. Rogoff, and Christoph Trebesch. 2020. "The Debt Pandemic." *Finance and Development* 57 (3): 12–6.

Chabert, Guillaume, Martin Cerisola, and Dalia Hakura, 2022. "Restructuring Debt of Poorer Nations Requires More Efficient Coordination. *IMF Blog*, April 7. https://blogs.imf.org/2022/04/07/restructuring-debt-of-poorer-nations-requires-more-efficient-coordination/.

Du, Wenxin, and Jesse Schreger. 2016. "Local Currency Sovereign Risk." *Journal of Finance* 71 (3): 1027–69.

Du, Wenxin, and Jesse Schreger. 2022. "CIP Deviations, the Dollar, and Frictions in International Capital Markets." In *Handbook of International Economics*, Vol. 6, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 147–97. Amsterdam: Elsevier.

Du, Wenxin, and Jesse Schreger. Forthcoming. "Sovereign Risk, Currency Risk, and Corporate Balance Sheets." *Review of Financial Studies.*

Eaton, Jonathan, and Mark Gersovitz, 1981. "Debt with Potential Repudiation: Theoretical and Empirical Analysis." *Review of Economic Studies* 48 (2): 289–309.

Fang, Xiang, Bryan Hardy, and Karen K. Lewis. 2022. "Who Holds Sovereign Debt and Why It Matters?" National Bureau of Economic Research Working Paper 30087.

FRED. *Federal Reserve Economic Data***.** St. Louis: Federal Reserve Bank of St. Louis (accessed August 17, 2022).

Horn, Sebastian, Carmen M. Reinhart and Christoph Trebesch. 2019. "China's Overseas Lending." Unpublished.

Ilzetzki, Ethan, Enrique G. Mendoza, and Carlos A. Végh. 2013. "How Big (Small?) Are Fiscal Multipliers?" *Journal of Monetary Economics* 60 (2): 239–54.

Ilzetzki, Ethan, Carmen M. Reinhart, and Kenneth S. Rogoff. 2020. "Will the Secular Decline in Exchange Rate and Inflation Volatility Survive Covid-19?" *Brookings Papers on Economic Activity* 46 (2): 279–332.

Ilzetzki, Ethan, Carmen M. Reinhart, and Kenneth S. Rogoff. 2022. "Rethinking Exchange Rate Regimes." In *Handbook of International Economics*, Vol. 6, edited by Gita Gopinath, Elhanan Helpman, and Kenneth S. Rogoff, 91–145. Amsterdam: Elsevier.

IMF. 2021a. *India: 2021 Article IV Consultation-Press Release; Staff Report; and Statement by the Executive Director for India*. Washington, DC: IMF Staff Country Reports.

IMF. 2021b. *Fiscal Monitor: Strengthening the Credibility of Public Finances*. Washington, DC: IMF.

IMF. 2022a. *World Economic Outlook: Update—Gloomy and More Uncertain*. Washington, DC: IMF.

IMF. 2022b. "Total reserves (includes gold, current US$)." World Bank. https://data.worldbank.org/indicator/FI.RES.TOTL.CD (accessed September 2, 2022).

Kose, Ayhan, Sergo Kurlat, Franziska Ohsnorge, and Naotaka Sugawara. 2017. "A Cross-Country Database of Fiscal Space." Unpublished.

Kose, Ayhan, Peter Nagle, Franziska Ohsnorge, and Naotaka Sugawara. 2021. *Global Waves of Debt: Causes and Consequences*. Washington, DC: World Bank.

Kose, M. Ayhan, Franziska L. Ohnsorge, Carmen M. Reinhart, and Kenneth S. Rogoff. 2022. "The Aftermath of Debt Surges." *Annual Review of Economics* 14 (1): 637–63.

Krueger, Anne O. 2002. *A New Approach to Sovereign Debt Restructuring*. Washington, DC: IMF.

Mauro, Paolo, and Jing Zhou. 2021. "r – g < 0: Can We Sleep More Soundly?" *IMF Economic Review* 69 (1): 197–229.

Maggiori, Matteo, Brent Neiman and Jesse Schreger. 2020. "International Currencies and Capital Allocation." *Journal of Political Economy* 128 (6): 2019–66.

Miranda-Agrippino, Silvia, and Hélène Rey. 2020. "U.S. Monetary Policy and the Global Financial Cycle." *Review of Economic Studies*, 87 (6): 2754–76.

Obstfeld, Maurice, and Kenneth Rogoff. 1995. "The Mirage of Fixed Exchange Rates." *Journal of Economic Perspectives* 9 (4): 73–96.

Obstfeld, Maurice and Kenneth Rogoff. 1996. *Foundations of International Macroeconomics*. Cambridge, MA: MIT Press.

Okonjo-Iweala, Ngozi, Tharman Shanmugaratnum, and Lawence Summers. 2021. "Rethinking Multilateralism for a Pandemic Era." *Finance and Development*. December, 6–9

Ostry, Jonathan D., Atish R. Ghosh and Rafael Espinoza, 2015. *IMF Staff Discussion Note: When Should Public Debt be Reduced?* Washington, DC: IMF.

Qian, Rong, Carmen M. Reinhart, and Kenneth S. Rogoff. 2011. "On Graduation from Default, Inflation and Banking Crises: Elusive or Illusion?" In *NBER Macroeconomics Annual*, Vol. 25, edited by Martin Eichenbaum, Erik Hurst, and Jonathan A. Parker, 1–36. Chicago: University of Chicago Press.

Reinhart, Carmen M., Kenneth S. Rogoff, and Miguel A. Savastano. 2003. "Debt Intolerance." *Brookings Papers on Economic Activity* 34 (1): 1–74.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. *This Time is Different: Eight Centuries of Financial Folly*. Princeton: Princeton University Press.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2011. "From Financial Crash to Debt Crisis." *American Economic Review* 101 (5): 1676–706.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2014. "Financial and Sovereign Debt Crises: Some Lessons Learned and Those Forgotten." In *Financial Crises: Causes, Consequences, and Policy Responses*, edited by Stijn Claessens, M. Ayhan Kose, Luc Laeven, and Fabián Valencia, 141–55. Washington, DC: IMF.

Reinhart, Carmen M., Vincent R. Reinhart, and Kenneth S. Rogoff. 2012. "Public Debt Overhangs: Advanced-Economy Episodes since 1800." Journal of Economic Perspectives 26 (3): 69–86.

Rogoff, Kenneth S. 1985. "The Optimal Degree of Commitment to an Intermediate Monetary Target." *Quarterly Journal of Economics* 100 (4): 1169–89.

Rogoff, Kenneth S. 2019. "The Case for a World Carbon Bank." *Project Syndicate*, July 8. https://www.

project-syndicate.org/commentary/world-carbon-bank-for-developing-countries-by-kenneth-rogoff-2019-07.

Rogoff, Kenneth S. 2020. "Falling Real Interest Rates, Rising Debt: A Free Lunch?" *The Journal of Policy Modelling* 42 (4): 778–90.

Rogoff, Kenneth S. 2022. "Replication data for: Emerging Market Sovereign Debt in the Aftermath of the Pandemic." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E178001V1.

Schlegl, Mattias, Christoph Trebesch, and Mark Wright. 2019. "The Seniority Structure of Sovereign Debt." Unpublished.

World Bank. 2020. *Global Financial Development Report 2019/2020: Bank Regulation and Supervision a Decade after the Global Financial Crisis*. Washington, DC: World Bank

World Bank. *International Debt Statistics*. Washington, DC : World Bank (accessed August 17, 2022).

World Bank. 2022. *World Development Report 2022: Finance for an Equitable Recovery*. Washington, DC: World Bank.

Yue, Vivian Z. 2010. "Sovereign Default and Debt Renegotiation." *Journal of International Economics* 80 (2): 176–87.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 167–192

Popular Personal Financial Advice versus the Professors

James J. Choi

Millions of people get financial advice from noneconomists. Robert Kiyosaki's book *Rich Dad Poor Dad* has sold 32 million copies since 1997 (Lisa 2021). Dave Ramsey's book *Total Money Makeover* has sold 1.5 million copies since 2013 (NPD 2020), and his website reports that his radio show attracts 18 million listeners per week. Authors like these may be more influential than economists are. Chopra (2021) finds that exposure to Dave Ramsey's radio show, which promotes high saving rates, reduces household retail spending tracked by the Nielsen Homescan panel by at least 5.4 percent.

What advice are popular authors giving, and how does it compare to the prescriptions of economists' normative models? In this article, I survey the advice given by the 50 most popular personal finance books (listed in Appendix Table 1) as ranked by the website Goodreads in May 2019. Three of the books contain no advice on the topics on which I focus here, resulting in a final sample of 47 books. Table 1 summarizes—with some oversimplification—the consensus advice of popular authors and the corresponding advice from benchmark academic theories. The upshot is that popular advice often deviates from economists' advice.

Understanding popular personal financial advice is interesting for several reasons. First, popular advice may help us interpret why the financial choices we observe empirically arise. It is plausible that some choices that depart from economists' theoretical predictions are being driven by the reasoning and

■ *James J. Choi is Professor of Finance at the Yale School of Management, New Haven, Connecticut. He is a Research Associate at the National Bureau of Economic Research, Cambridge, Massachusetts.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.167.

Table 1
Summary of Consensus Popular Advice and Benchmark Academic Advice

Topic	*Consensus popular advice*	*Benchmark academic advice*
Saving	Save 10–15 percent of income regardless of age and circumstances during working life. Don't annuitize. Spend to keep real level of wealth roughly constant in retirement. Divide savings into mental accounts devoted to different goals.	Smooth consumption over time. Low or negative savings rates when young, high savings rate in midlife. Fully annuitize wealth in retirement. If not annuitized, negative savings rate in retirement. All wealth is fungible.
Portfolio equity share	Hold money that might be spent in short term entirely in cash. Money that won't be spent in short term may be invested in equities. Equity share should be hump-shaped with respect to age.	Invest money that will fund near-term consumption more conservatively than money that will fund consumption far in the future. Equity share should be hump-shaped with respect to age. Equity share should depend on how quickly marginal utility diminishes, how much stock returns covary with marginal utility.
Dividends	High dividends are attractive.	High dividends are unattractive.
Equity styles	Value stocks and small stocks are attractive.	Value stocks and small stocks may or may not be attractive.
International diversification	Hold international stocks, but far less than in proportion to their global market cap weight.	Hold international stocks in proportion to their global market cap weight.
Active mutual fund management	Invest only in passive index funds.	Invest only in passive index funds.
Non-mortgage debt paydown	Either prioritize paying highest-interest debt or lowest-balance debt. Co-holding low-interest assets and high-interest debt may be a good idea.	Prioritize paying highest-interest debt. Do not co-hold low-interest assets and high-interest debt.
Mortgage choices	Choose a fixed-rate mortgage.	Choose an adjustable-rate mortgage unless interest rates are low.

recommendations described by popular authors.[1] Second, popular advice might contain valuable normative insights that economists have overlooked.[2] Third, even

[1] Respondents to surveys conducted by Choi and Robertson (2020) and Bender et al. (forthcoming) rate "advice from a book or an article I read, or somebody on TV, radio, or the internet" as one of the least important factors for determining their portfolio's equity share. However, individuals' choices could be driven by lay reasoning that is *reflected* in popular authors' writings, even if choices are not causally affected by the authors.

[2] An instance where economists overlooked the wisdom in popular advice before accepting it is found in Canner, Mankiw, and Weil (1997) and the response by Brennan and Xia (2000), Campbell and Viceira (2001), and Wachter (2003) regarding how long-term bond holdings should respond to risk aversion.

if popular advice is not exactly optimal, it may be second best in a way that illuminates the constraints faced by individuals. For example, one common reason for why popular advice might deviate from economists' advice is that popular advice tries to take into account the limited willpower individuals have to stick to a financial plan.

I begin by surveying advice on choosing savings rates over the lifecycle as well as the advisability of being a wealthy hand-to-mouth consumer who has substantial illiquid assets such as housing but almost no liquid wealth. Next, I cover advice on asset allocation—the fraction of your financial portfolio that should be invested in stocks, securities that pay dividends or interest, equity styles (colloquially known as "smart beta"), international diversification, and active versus passive mutual fund investment. The following section summarizes recommended strategies for managing non-mortgage debt—in particular, which debts to prioritize repaying and whether simultaneously holding low-interest-earning assets and high-interest debt is a good idea. The final section covers advice about mortgage choices—fixed-rate versus adjustable rate, the size of the down payment, maturity, paying principal ahead of schedule, and refinancing.

Consumption and Savings

Economic Theory on Savings over the Lifecycle

Economists think about optimal savings rates in a way that is probably counterintuitive to the layperson. Economic theory targets an optimal *consumption* rate each period. The optimal savings rate is whatever the difference happens to be between income and optimal consumption. In the standard lifecycle/permanent income hypothesis model with neither uncertainty nor borrowing constraints, individuals choose a consumption growth rate that trades off smoothing marginal utility (and hence the level of consumption) over time against the rate of time preference (that is, the preference for earlier gratification) and the financial return from saving. The desired consumption growth rate and the lifetime budget constraint jointly determine the starting level of consumption. Individuals with constant relative risk aversion utility consume the same amount every period if the rate of time preference equals the interest rate. Because income tends to be hump-shaped with respect to age, savings rates should on average be low or negative early in life, high in midlife, and negative during retirement. From this perspective, the common policy of a default retirement savings plan contribution rate that does not depend on age is suboptimal.

Carroll (1997) shows that if individuals are sufficiently impatient relative to the expected growth rate of their income and labor income is risky, they will be "buffer-stock savers"—that is, aiming to accumulate only a few months' worth of income in assets to insulate against income fluctuations. Their average savings rate is close to zero after the target asset level is reached, adjusting mainly to keep asset balances near the target level. To guard against the possibility of a catastrophically

low labor income realization, buffer-stock savers do not borrow. Carroll (1997) and Gourinchas and Parker (2002) estimate that the typical household is a buffer-stock saver until midlife, at which point it switches over to accumulating greater sums in order to prepare for retirement.

Popular Authors on Saving during Working Life

In contrast to the emphasis on smoothing consumption in economic theory, popular authors advise smoothing savings rates, which is also the default option for a typical retirement savings plan. Dacyczyn (1998, p. 548) is the only popular author who makes the economist-like observation that "the semiresponsible admonition to save 10 percent of your income essentially endorses the constant contracting and expanding of family expenditures. But surprisingly, life is easier and more enjoyable if spending always stays, on average, at a modest level." The advice of Robbins (2014, p. 58) is more typical in running counter to consumption smoothing:

> Whatever that [savings percentage] number is, you've got to stick to it. In good times and bad. No matter what. Why? Because the laws of compounding punish even one missed contribution. Don't think of it in terms of what you can afford to set aside—that's a sure way to sell yourself short. And don't put yourself in a position where you can suspend (or even invade) your savings if your income slows to a trickle some months and money is tight.

Chilton (2011, pp. 95–6) explicitly rejects consumption smoothing:

> Strangely, a few economists and mathematicians have been pushing the idea of intentionally not saving in your early working years because your income is low and your starting-out-in-life costs are high. They advocate ramping up efforts big time in your middle years . . . Do not heed that advice . . . it seldom works in the living room. First, costs have a funny way of never stabilizing. Second, most people aren't going to be able to transition from setting aside nothing to being supersavers at the flip of a switch. Psychologically, that's just not realistic. Finally, I can't get the numbers to work anyway.

These passages illustrate two key motivations for the popular advice to smooth savings rates: the usefulness of establishing saving consistently as a discipline, and the power of compound interest. The discipline argument is almost always missing from economic models of optimal saving—a potentially important oversight. Economic models do make the opportunity cost of consumption a key driver of savings recommendations, but in such models, it could result in either a negative or positive optimal current savings rate. In contrast, for popular authors, compound interest implies that you should be saving a positive amount now. Of the 45 books that offer some sort of savings advice, 32 stress the importance of starting to save immediately, and 31 regale the reader about the power of compound

interest.[3] Contrary to the advice of a lifecycle/permanent income hypothesis model that the young should often have negative net worth, 28 books mention the need for everybody to prioritize building an emergency savings buffer of between $1,000 and two years of income. Contrary to the buffer stock model's recommendation, popular authors advise you to continue saving at a high rate even after an adequate emergency savings fund has been established.

Popular authors tend to champion simple savings rules that have the virtue of requiring minimal computation. Twenty-one books recommend a positive savings rate that does not vary by age. Ten to 15 percent of income is a range that encompasses most of the recommendations. Four books recommend 20 percent or a range that includes 20 percent, and two recommend 50 percent or more on the premise that you should have the financial freedom to be able to walk away from an undesirable job early in life. Only one of the above 21 recommendations adjusts for the amount you have already saved. Nine books advise starting with a target for wealth at retirement in order to compute the necessary savings rate; many of these books compute the constant dollar savings flow per period needed to achieve the goal. Only four books recommend taking Social Security benefits into account when choosing a savings rate, despite Social Security replacing 64 percent of final working-life earnings for the median new beneficiary aged 64–66 in 2005 (Biggs and Springstead 2008).

If your employer matches employee contributions to a 401(k) retirement savings plan, 11 books recommend contributing enough to earn the maximum possible match. Failing to do so is "like walking away from free money" (Kobliner 2017, p. 136). Nobody recommends altering your savings rate in response to how generous the match is. As Ramsey (2003, p. 158) puts it, "if your employer matches . . . that amount is gravy." The maximum allowable annual contribution to your 401(k) or Individual Retirement Account is another focal amount recommended by six books.

Fifteen books recommend increasing your savings rate over time. If income is rising over time, this strategy is consistent with consumption smoothing. Indeed, four books recommend diverting some of future salary increases to savings rate increases. However, three books recommend increasing your savings rate by 1 percent of income per month over the next few months—faster than plausible income growth for most people—on the theory that you can acclimate to a higher savings rate over time. Eight books say that a lower consumption level becomes easier to tolerate with the passage of time.

Orman (2007) alone says that it is reasonable to carry revolving balances on one's credit card in one's younger years in anticipation of higher future income; she advises that no more than 1 percent of current pretax annual income be charged each month to the credit card, and these charges should only be for absolute necessities. Thirty-one books warn against borrowing on credit cards, usually in strong terms such as, "Credit card debt is never good" (Kobliner 2017, p. 33). Ramsey

[3]My classifications of each book's advice, along with the relevant textual excerpts, are available in the Online Appendix.

(2013, p. 126) gives the most extreme advice against high-frequency consumption smoothing using debt, writing, "The worst time to borrow is when times are bad," on the grounds that the debt payments will be burdensome if income doesn't recover. Eighteen books give some variant of the advice that debt can be good when used to fund investments in things that appreciate, such as houses and human capital, but is bad otherwise. Seven books advise against student loans.

Mental Accounting

Standard economic theory does not earmark portions of household savings for specific purposes; money is fungible. In contrast, 17 books advocate subdividing wealth into mental accounts devoted to different goals. Malkiel (2019, p. 358) writes, "A specific need must be funded with specific assets dedicated to that need." Commonly mentioned mental accounts are a fund for emergencies, a retirement savings fund, a fund for major purchases such as a house or a car, and a fund for children's college tuition. The previously mentioned recommended savings rates are usually for the retirement and/or emergency funds alone; saving for other expenses is to be done in addition to the baseline savings level.

Mental accounting can create various choice pathologies due to not considering one's overall combined financial position (Thaler 1985, 1999). However, using mental accounting when choosing savings rates has some advantages. Karlan et al. (2016) argue that mental accounting increases motivation to save by making salient the link between today's saving and specific future expenditures. Popular author Kobliner (2017, p. 28) asserts, "Research . . . suggests that labeling a savings account with a goal . . . actually results in people adding even more money to their savings pot." Mental accounting is also a straightforward way to compute how to adjust spending in response to foreseeable changes in the utility from expenditure. Economic models typically assume that the individual's utility function is the same every period and that goods are infinitely divisible, which makes optimal expenditure levels smooth across time. In reality, expenditure is much more valuable in certain periods—say, when you are getting married, moving to a new city, or sending a child to college—and needs to be lumpy in order to purchase durable goods, which limits the practical appeal of exactly following the savings recommendations from standard economic models. Funding a mental account devoted to a predictable large expenditure in addition to your baseline saving has the desirable effect of decreasing spending before the period when expenditure is unusually valuable. Such a practice also helps you monitor whether you are on track to have enough money to fund the expenditure when the time comes.

Reframing Saving: Pay Yourself First

One trope in popular finance recurs so often that it is worth mentioning, as its prevalence across books and decades suggests that it is effective. Economists conceive of current consumption as the carrier of utility and saving as a current sacrifice. Clason (1926) was the first to reframe current saving as the carrier of utility—a payment to yourself—while current spending is reframed as a payment

to others and hence a loss to yourself. His rule that you should "pay yourself first" appears in 16 books.

The rule prescribes that when you receive income, a predetermined fraction should be sent to a separate account at once—automatically, advise many modern books—and this money is not to be touched. "The secret . . . is that you can't spend what you don't see" (Bach 2004, p. 20). The remaining money can be freely spent without careful budgeting. This advice is frequently accompanied by Clason's (1926, p. 19) statement that after increasing one's savings rate by 10 percentage points, "strange as it may seem, I was no shorter of funds than before. I noticed little difference as I managed to get along without it." The notion that a significant amount of the money we spend brings us almost no marginal utility—making additional saving painless—is endorsed by 18 books.

Wealthy Hand-to-Mouth Status

Kaplan, Violante, and Weidner (2014) document that about 20 percent of US households are "wealthy hand-to-mouth," in that they have positive illiquid assets, such as housing and retirement account balances, but almost no liquid assets. Kaplan, Violante, and Weidner (2014) and Kaplan and Violante (2014) find that such a portfolio composition can be rationalized despite the resulting inability to cushion consumption from income shocks if illiquid assets have extremely high risk-adjusted returns. Angeletos et al. (2001) instead interpret this pattern as the result of households storing wealth in illiquid forms to protect it from their lack of self-control.

If illiquid assets do yield such high returns that it is worth living paycheck to paycheck, then we might expect to see evidence of such a belief in popular advice. I have already mentioned that 28 books explicitly state the need for everybody to prioritize building emergency savings, which is contrary to being a wealthy hand-to-mouth household. Fourteen books say that a house is not a great financial investment. Of the seven books that say that a house is a great investment, six recommend building emergency savings of at least three months' income/expenses, and two warn against becoming a wealthy hand-to-mouth household in order to buy a more expensive house.

Thus, popular authors do not seem to believe that it is advisable to become a hand-to-mouth consumer in order to invest in housing. However, going without a liquid asset buffer in order to take advantage of 401(k) matching contributions—a high, instantaneous, and risk-free return on investment—is viewed more sympathetically. Many of the 11 books that recommend contributing enough to receive the maximum possible 401(k) match do not give advice on how to trade off 401(k) contributions against emergency savings. But three books do recommend prioritizing 401(k) contributions over building an emergency cash buffer.

Spending in Retirement

Fifteen books give advice on spending in retirement. Two books advise planning on lower spending in retirement than during working life, whereas two books advise keeping spending constant across the retirement threshold. Therefore, little light

is shed on the optimality of the empirically observed drop in spending that occurs upon retirement (Bernheim, Skinner, and Weinberg 2001; Aguiar and Hurst 2005).

One book advises spending 3 percent of your financial wealth per year in retirement, 7 books advise 4 percent, 1 advises 5 percent, 1 advises 6.7 percent, and 2 (both by Dave Ramsey) advise 8 percent on the theory that nominal investment returns will be 12 percent and the inflation rate will be 4 percent.[4] Five books explicitly tie their recommended withdrawal rate to be at or below a stated expected real portfolio return, implying that preserving the real level of capital is the goal, rather than spending down wealth as the lifecycle model implies.

One reason to avoid decumulating wealth in retirement is to ensure that you don't outlive your savings. The classic model of Yaari (1965) advocates fully annuitizing wealth to eliminate this longevity risk, but only four books recommend buying life annuities. No book explicitly recommends against life annuities, but we might gain some insight into why households buy so few annuities from some of the drawbacks listed or refuted by the books that do encourage annuitization: the risk of early death, the loss of control over one's money, low current interest rates, and the fact that most annuities lack inflation protection.

Asset Allocation

Popular Authors on the Equity Share of Portfolio

Investment time horizon is of paramount concern in popular financial advice. Thirty-one of the 45 books that offer some form of asset allocation advice assert that stocks become less risky, and hence more attractive, as the holding period increases. Twenty of these books justify this argument by pointing to the fact that historically, stocks were less likely to underperform fixed-income assets or to have a negative cumulative return as the holding period increased. Twelve books say that stock returns mean-revert; a common saying is that stocks are "on sale" after a large price decline. Seven books argue that because the economy will grow in the long run, stocks are likely to be profitable investments in the long run.

The perception that stock market risk decreases with holding length leads to popular authors recommending that stock allocations increase with investment horizon. Money is often bucketed by when it will be needed, and a different investment allocation is recommended for each bucket. Twenty-nine books say that money that might be spent soon should be held entirely in cash. In particular, emergency savings should be held in cash, usually in a bank account. Many authors also recommend that non-emergency savings that will be needed in the near-term should be held in cash or fixed income, where "near-term" is defined as one year (one book),

[4] Some books advise spending X percent of your wealth during your first year of retirement, and to then grow that spending amount at the inflation rate. Other books are vague about how spending should adjust over time in response to changes in one's portfolio value.

one or two years (one book), one to three years (one book), five years (seven books), two to seven years (one book), or even as long as ten years (two books).

Longer-term money such as retirement savings is to be invested more heavily in equities, although 14 books warn against allocating 100 percent to equities because such a portfolio is too risky and lacks diversification across asset classes. Twenty-six books recommend that the asset allocation of long-term money become more conservative with age; nine cite a variant of the "portfolio percent in stocks should be 100 minus your age" rule. Four books recommend that *any* money not needed in the near-term be invested in stocks. These rules create a hump-shaped pattern of portfolio equity share with respect to age. The young have little surplus money that is not needed in the near term, so most of their financial assets will be in fixed income assets. The middle-aged have accumulated more savings, so a greater portion of their financial assets are deemed suitable for equity investment because they are not needed for short-term spending. Older individuals have a lower equity share because of the growing conservatism prescribed for their long-term money and because more of their money will be needed in the near-term to fund retirement consumption.

The inflation rate is mentioned by 11 books as a reference return that is important for one's longer-term money to exceed. For example, Ferri (2010, p. 94) recommends, "Each asset class to be held in a portfolio for the long term should be expected to earn a return greater than the inflation rate." Following such a decision rule implies that risk-taking will increase when real interest rates become negative.

Interpreting Empirical Evidence on Lifecycle Portfolio Allocation

Popular advice offers guidance on how economists should interpret empirical patterns in lifecycle asset allocation. Ameriks and Zeldes (2004) point out that even with perfect panel data, it is impossible to identify how asset allocation changes with age without imposing a strong assumption about the shape of age, birth cohort, and calendar time effects, since these three variables are perfectly collinear—a person's age equals the calendar year minus the person's birth year. Depending on the identifying assumption used, they find that the percent of portfolio allocated to equities is either strongly rising or hump-shaped with respect to age.

In contrast to this econometric ambiguity, none of the books in the sample recommends that one's stock allocation should be everywhere increasing with age. The weight of popular recommendations suggests that individuals do not have portfolio rules in mind that everywhere increase equity share with age.

A New Explanation for Stock Market Non-Participation

Popular advice suggests an explanation for stock market non-participation that is absent from the academic literature. Only half of US households hold any stock either directly or indirectly via mutual funds or pension funds (Guiso and Sodini 2013). Non-participation is a puzzle for economic theory because under expected utility preferences, everybody should hold at least a small amount of stock provided that their non-stock income is not too positively correlated with

stock returns (Haliassos and Bertaut 1995).[5] The fact that stock market participation rises with wealth has caused the existence of fixed costs of participation to become a leading candidate explanation for non-participation (Vissing-Jørgensen 2003). Because the expected benefit of investing a minimal amount in the stock market must be small, a fixed cost will deter stock market investment among those with little wealth.

But if many people believe that any money that may be spent in the near term should not be invested in stocks, then low stock market participation rates that rise with wealth are a natural outcome even in the absence of fixed costs.

Economists on the Equity Share of Portfolio

Economic models also recommend equity allocations that are hump-shaped with age, but for somewhat different reasons than those found in popular books.

Unlike many popular authors, economists commonly think of investment risk not in terms of the probability of the binary outcome that stocks will outperform bonds or have a positive return, but in terms of the variance of returns. This measure recognizes that the magnitude by which stocks outperform or underperform the safe asset should matter. In the benchmark case where stock returns are identically and independently distributed, the expected outperformance of stocks over bonds increases with investment horizon, but this is offset by the fact that the variance of cumulative stock returns also increases. Samuelson (1969) and Merton (1969) prove that if the investor has constant relative risk aversion utility and no labor income, the optimal allocation to the stock market does not vary with the investment horizon.

The story is different if stock returns are negatively autocorrelated, which causes the annualized variance of cumulative log stock returns to decrease with investment horizon. Campbell, Lo, and MacKinlay (1997, p. 80) survey the academic literature and conclude that "there is little evidence of mean reversion of long-horizon returns" of the form popular authors describe, where low recent stock market returns unconditionally predict high future stock market returns. However, low stock market returns that lower the price-dividend ratio because they are not accompanied by a contemporaneous drop in dividends *do* forecast high future returns (Cochrane 2009, pp. 422–4). Barberis (2000) finds that if one takes into account today's price-dividend ratio and expectations of how it will evolve in the future, the *conditional* annualized variance of cumulative log stock market returns declines with horizon. Conditional variances are the theoretically relevant object for the investment decision, since they represent the uncertainty that the investor faces after considering her information set. My sense is that most financial economists believe that conditional annualized stock market risk decreases with horizon. However, Pástor and Stambaugh (2012) argue that an expansive view of parameter uncertainty implies that conditional annualized stock market risk *increases* with horizon.

[5] Any strictly increasing differentiable utility function is locally linear, so agents with such utility functions should be risk-neutral with respect to a small risk (Rabin 2000).

If conditional annualized return variance decreases with investment horizon, then economic theory generally recommends that long-horizon investors whose coefficient of relative risk aversion is above one (the empirically relevant case) hold more stock than short-horizon investors (Barberis 2000). Wachter (2002) finds that when stock returns are mean-reverting and perfectly correlated with a return predictor (for example, the dividend-price ratio), investors will optimally break up their portfolios into subaccounts for funding each consumption event. Money that is intended to be spent further in the future should be invested more aggressively. This bucketing strategy is akin to the approach popular authors recommend, although money intended for near-term use should not generally be invested entirely in cash.

Standard theory also departs from popular advice in not regarding the *level* of the risk-free interest rate as relevant for portfolio allocation, but only the *difference* between expected risky asset returns and the risk-free interest rate. In other words, one should not become more prone to reach for yield in low-interest-rate environments if risk premia remain unchanged, although Campbell and Sigalov (2022) present a model where a requirement to consume the expected return of the portfolio each period (a "sustainable spending" constraint) could justify such behavior.

A rationale for portfolio equity shares that decline with age that does not depend upon return mean reversion lies with human capital (Bodie, Merton, and Samuelson 1992). If labor income is like a bond interest payment that is relatively uncorrelated with stock returns, then a young person has an implicit fixed-income position whose value is usually enormous relative to that person's financial assets. As the person ages, the present value of future labor income declines because there are fewer wage payments remaining. To offset the decline in implicit fixed-income holdings, the financial portfolio should hold more fixed income over time. Labor supply flexibility also increases the capacity to bear risk in one's financial portfolio, because a low investment return can be mitigated by working more. If the young have more labor supply flexibility than the old, then this is another reason for the young to hold a greater share of stocks in their portfolios than the old.

Cocco, Gomes, and Maenhout (2005) find that in a lifecycle model with fixed labor supply, human capital causes those whose risk aversion is at the upper boundary of what is usually thought to be plausible to allocate 100 percent of their portfolios to equities for much of working life before gradually reducing that equity share as retirement approaches. Adding the possibility of disastrous transitory labor income shocks makes the young invest more conservatively than the middle-aged due to the increase in background risk. A fixed cost of stock market participation also deters the young, who have low asset holdings, from investing in stocks (Gomes and Michaelides 2005).

Missing Drivers of Equity Share in Popular Advice

Despite the importance of human capital, only eight popular books mention it as a relevant consideration for lifecycle asset allocation. All eight indicate that higher future wage earnings increase optimal risk-taking in the financial portfolio,

but none explicitly mention labor supply flexibility, instead writing things like, "[the young] can use wages to cover any losses from increased risk" (Malkiel 2019, p. 344).

Two concepts that are foundational for portfolio choice theory are rarely if ever mentioned by popular authors: diminishing marginal utility of consumption and return covariance with marginal utility.

Economists conceive of risk aversion as being driven by the speed at which marginal utility diminishes as consumption increases. Swiftly diminishing marginal utility means that the upside potential of a gamble is not so valuable, making gambles less attractive. Only five books suggest that diminishing marginal utility should be a determinant of one's portfolio equity share. All five give the impression that diminishing marginal utility is relevant only after one achieves or is close to achieving one's target wealth level. For example, Ferri (2010, p. 285) writes, "You only take the amount of risk that you need to accomplish a financial objective . . . There is no need to invest at your peak risk tolerance level once you have accumulated enough assets to easily reach your investment objectives with lower risk." Three books argue that investors do not know their own risk tolerance—which they define as the ability to not sell your stock in a bear market, rather than the speed with which marginal utility diminishes—until they have lived through a major market decline. Thus, they recommend that younger investors scale back the risk of their portfolios until they have gained such experience.

Another fundamental driver of asset allocation in economic models is the covariance of each asset's return with marginal utility: investors should be more reluctant to hold assets that tend to deliver low returns when an extra dollar is especially valuable. This means that in equilibrium, such assets will have low prices and hence high average returns. (The previously discussed variance of returns matters for portfolio choice only to the extent that it is ultimately tied to this covariance.) The celebrated equity premium puzzle (Mehra and Prescott 1985) is the observation that stocks' returns don't seem to have a large covariance with marginal utility, as measured by aggregate consumption growth, and yet their average returns are high, making stocks anomalously good deals. It is striking that *none* of the popular advice books mention period-by-period covariance with marginal utility as a consideration for asset allocation. This suggests that consumption-based asset pricing models, which seek to explain expected returns using covariances of returns with consumption, may fail because people simply don't make portfolio decisions with such covariances in mind.

The closest any author comes to this notion is a concern, mentioned by 11 books, that one might be forced to sell prematurely at a loss. The act of selling plays a central role. Orman (2012, p. 246) writes, "If you don't have the time to leave this money sitting there, it is possible that when you do need to take it out, that need will arise at the worst possible time . . . One year later, you find the house you want and make the offer, which is accepted—on April 14, 2000, a day the market goes down considerably, and the day you had decided to sell. You will most likely take out far less than you initially put in. If you could have just waited—but you could not, for you needed the money to buy your home." Notice that this concern would apply

even to risky assets whose returns are uncorrelated (in expectation) with marginal utility, and that it would not apply if other assets were sold to finance expenses while the underwater asset were held. As Robin and Dominguez (2018, p. 292) write, in the minds of popular authors, "The only days you care about an investment's value are the day you buy it and the day you sell it."

Dividends and Interest

Miller and Modigliani (1961) prove that in a frictionless market with no taxes, a firm's payout policy is irrelevant for its valuation. The intuition is that any investor who desires a certain amount of cash from an investment can generate it by selling shares, instead of relying on a dividend. In the real world, dividends and interest are disadvantaged relative to capital gains in the US tax code, which makes the prevalence of dividends a puzzle (Baker and Weigand 2015).

Nine of the books in our sample reject the dividend irrelevance theorem, and no book recommends eschewing dividends or interest for tax reasons. Multiple books refer to the need for "income," particularly when the investor is older, for which bonds are the preferred source. Malkiel (2019) recommends coping with a low-interest-rate environment by holding relatively stable dividend-paying stocks in place of what would be bond holdings in normal times. Kiyosaki (2012) dismisses the relevance of capital gains, arguing that cashflow from the investment is the only relevant factor. Relatedly, Ferri (2010, p. 30) writes that commodities have lower expected returns than stocks because they "pay no interest, have no earnings, and pay no dividends," which seems to be an expression of the fallacy that dividend payments do not come at the expense of capital gains (Hartzmark and Solomon 2019). Lynch (1989, p. 205) argues that "the presence of the dividend can keep the stock price from falling as far" because "if investors are sure that the high [dividend] yield will hold up, they'll buy the stock just for that." But inconsistent with this assertion, from July 1927 to June 2022, a value-weighted portfolio of all non-dividend-paying stocks has more positively skewed monthly returns than dividend-paying stocks in the bottom 30 percent or middle 40 percent of the positive dividend yield distribution.[6]

Equity Styles

Stocks with certain characteristics—or styles—have historically had higher average returns than stocks with the opposite characteristics. For example, value stocks (which have low prices relative to their current fundamentals such as book equity, dividends, or earnings) have had higher returns than growth stocks, and stocks of "small-cap" companies with a relatively small market capitalization have had higher returns than stocks of "large-cap" companies with a large market capitalization (Fama and French 1992). These average return differences do not appear to be compensation for bearing the classical measure of risk, "market beta" (the

[6]Value-weighted portfolio returns are obtained from French (2022). Stocks are sorted by dividend yield as of each June-end.

responsiveness of an asset's return to the aggregate stock market's return, which captures how an asset affects the variance of a well-diversified portfolio's return). This empirical finding has led to an active academic debate about whether average style returns are due to mispricing or rational compensation for risk that is not captured by market beta.

Twenty-six books offer a recommendation about equity style tilts. Eight books recommend tilting one's portfolio toward value stocks, while one book recommends tilting toward growth stocks. Fourteen books recommend tilting toward small stocks. The advice to diversify across opposing styles (and underweight stocks in the middle of the style dimension) is common, which weakens style tilts. Eight books recommend holding both growth and value stocks, while thirteen books recommend holding both large and small stocks in a way that creates a small-cap tilt.

Many fewer books say that these tilts could entail taking on more risk, which suggests that most authors think that their recommended tilts generate superior risk-adjusted returns. Only three books mention the possibility that value stocks are riskier than growth stocks. Ferri (2010, p. 91) writes that "growth stocks tend to perform well in a recession and early recovery, while value stocks tend to do best well into a recovery and at economic peaks." This appears to be untrue in the data; from July 1926 to June 2022, the Fama and French (1993) value-minus-growth factor (called HML) has an average monthly return of 0.37 percent during US recessions plus the first year of recovery, and 0.35 percent otherwise (French 2022b). Bernstein (2010, p. 120) writes: "Fama and French . . . insist that the higher return of value stocks reflects the fact that these companies . . . are weaker and thus more vulnerable in hard times . . ." but notes that "growth stocks demonstrate their own peculiar risks" because "from time to time, the public becomes overly enthusiastic about the prospects for companies at the leading edge of the era's technology." Bernstein (2017, p. 124) also warns about value stocks' risks for particular investors: "Employees of cyclical 'value' companies should be particularly wary of value portfolios, as in the event of a severe recession both their job prospects and their portfolios will suffer disproportionately." In contrast, six books say that growth stocks are riskier than value stocks. Four books say that small stocks are riskier than large stocks, and only one book says the opposite.

International Diversification

In a frictionless market with homogeneous investors, every investor should hold each country's securities in proportion to its market capitalization. In practice, investors heavily overweight the securities of their own country, foregoing significant diversification benefits (French and Poterba 1991).

Twenty-six books have something to say about international equity investment. Only two books recommend not diversifying internationally at all. The remainder recommend holding international stocks, but of those that give specific portfolio percentages, all recommend allocations that are below the 59 percent of global stock market capitalization that non-US stocks constitute as of 2021 (SIFMA

2021). The average recommended weight is 27 percent of equity holdings, with the range being from 12.5 percent to 50 percent.

Most books give no reasoning for why they underweight international stocks. Seven books say that international stocks are riskier than US stocks, citing higher return volatility, currency risk, lower liquidity, subpar accounting and financial transparency standards, and political instability. Bogle (1999, 2017) notes that a significant portion of the revenue and profits of S&P 500 companies comes from other nations, so US stocks already provide international exposure, and Collins (2016) writes that increasing cross-border market integration has reduced the diversification benefits of holding foreign stocks. Bogle (1999, 2017) and Collins (2016) also argue that the United States is the most attractive market to invest in because its economy will experience the strongest future growth. Bernstein (2017, p. 79) says that one's international stock exposure should be limited by how tolerable one finds it when one's portfolio "often temporarily underperforms everybody else's," given that one will be surrounded by other investors whose portfolios are home-biased.

Coeurdacier and Rey (2013) survey the academic literature on home bias. Many papers rationalize home bias by creating models where domestic equities are a hedge against the risk in income that can't be traded in financial markets (such as wages)—a motive that is not mentioned in any popular book. Conversely, the motives for underweighting international stocks that do appear in popular books tend to be rejected by economists. The foreign trading costs and perceived foreign risk due to low information quality necessary to rationalize the observed level of home bias are too large to be plausible (French and Poterba 1991). Currency risk can be hedged away at a cost that is negligible in major currencies (Perold and Schulman 1988). The correlation of multinationals' stock returns with their domestic stock market is very high, limiting the international diversification benefit obtained by buying the multinational stocks of one's own country (Lewis 1999). Bekaert, Hodrick, and Zhang (2009) and Christoffersen et al. (2012) find that trends in cross-market stock market return correlations have not eliminated the benefits of international diversification. Finally, a security's expected return equals its discount rate, regardless of its expected cashflow growth. Therefore, the perceived strength of the US economy is not a reason to overweight it if the market efficiently prices this strength.

Active versus Passive Mutual Fund Management

The average actively managed US equity mutual fund that tries to beat the market's return underperforms the average passive fund that tries to match the market's return by 0.67 percent per year (French 2008). Thus, mainstream economic advice is to avoid active management. Nevertheless, 60 percent of mutual fund and exchange-traded fund (ETF) assets in 2020 are invested in actively managed funds (Investment Company Institute 2021).

Popular authors largely advise investing in passive index funds. Twenty-four books recommend indexing. Only seven books recommend active management.

One of these books is by Peter Lynch, whose advice is not surprising given that he made his fortune as an active mutual fund manager. Authors that recommend active management generally recommend picking funds based on their past performance. Empirically, money flows into mutual funds strongly chase past returns (Chevalier and Ellison 1997), but evidence that performance persists is weak (Carhart 1997; Choi and Zhao 2021).

The agreement between popular advice and economists' advice may stem from the fact that the statistics on average performance and performance persistence are straightforward to calculate, easy to understand, and widely publicized.

Non-Mortgage Debt Management

Twenty-three books give advice on how to pay down non-mortgage debt, focusing predominantly on credit card debt.

Prioritizing Which Debt to Pay

For economists, a very basic principle of optimal debt repayment is to prioritize paying down the debt charging the highest interest rate.[7] In practice, households often do not follow this principle (Gathergood et al. 2019).

Surprisingly, ten books recommend *not* prioritizing one's highest-interest debt, versus ten books that endorse prioritizing one's highest-interest debt. Nine books endorse some variant of the "debt snowball" method, which is famously associated with Dave Ramsey. The debt snowball prioritizes paying off the smallest-balance debt first, while making the minimum required payment on the others. When the smallest-balance debt is paid off, the money that was being applied toward it now goes towards paying off the next-smallest-balance debt (in addition to the minimum payment on this next debt), and so on until all debts are paid off. Ramsey (2011, p. 100) writes, "People sometimes say, 'But Dave, doesn't it make more sense mathematically to pay off the highest interest rates first?' Maybe. But if you were doing math, you wouldn't have credit card debt, would you? This is about behavior modification. You need some quick wins or you will lose steam and get discouraged . . . every time you cross a debt off the list, you get more energy and momentum . . ." With a similar eye toward motivation, two books recommend prioritizing the debt that bothers you the most, regardless of its interest rate.[8]

[7] There are some caveats to this principle if defaulting on some debts is a significant possibility. For example, if a high-interest debt is easier to discharge in bankruptcy than a low-interest debt, it may make sense to deprioritize the former. If a low-interest debt is collateralized with an asset such as a house or a car that can be seized in default, it may be optimal to prioritize paying this debt while defaulting on an uncollateralized debt with a higher interest rate.

[8] Three of the above books advise prioritizing either the highest-interest debt or the lowest-balance debt. One of the books advises prioritizing either the lowest-balance debt or the debt that bothers you the most.

Co-holding of High-Interest Debt and Low-Interest Assets

Twelve books say that in order to pay off one's debt, it is important to establish a firm rule that one will not borrow anything more. For example, Warren and Tyagi (2005, p. 144) write, "This is the moment to look at yourself in the mirror and say out loud, '*The debt stops here.*' Every morning tell yourself, 'I will not take on more debt today.'"

The existence of this rule potentially gives some insight into the co-holding puzzle—the fact that 33 percent of households who are paying high interest rates on their credit card debt simultaneously hold at least one month of income in liquid assets earning low interest rates (Gross and Souleles 2002). The interest rate spread is large; in May 2022, the average rate on credit card accounts being charged interest was 16.65 percent, while the average savings account interest rate was 0.06 percent (Board of Governors of the Federal Reserve System (US); Federal Deposit Insurance Corporation). Economists have tried to rationalize co-holding by appealing to the fact that some expenses must be paid by cash or check (Telyukova and Wright 2008), strategic maneuvering in advance of bankruptcy (Lehnert and Maki 2007), attempts to limit household spending by reducing unused credit capacity (Bertaut, Haliassos, and Reiter 2009), and insuring against the risk that one's credit limit will be reduced (Fulford 2015).

Fourteen books endorse co-holding, but for very different reasons. Only one of them mentions in passing any justification found in the academic literature—the risk of a credit limit reduction.[9] Among the eleven books that say something against co-holding, eight nonetheless recommend some positive amount of co-holding. The most frequently cited justification for co-holding (mentioned by seven books) is that it prevents borrowing additional amounts. Warren and Tyagi (2005, p. 147) write, "This [emergency savings buffer] is the money that will keep you from sliding back into the credit card trap when something goes wrong." Ramsey (2013, p. 100) says that he used to recommend devoting all assets to paying down debt, but "I discovered that people would stop their whole Total Money Makeover because of an emergency—they felt guilty that they had to stop debt-reducing to survive . . . If you use debt after swearing off it, you lose the momentum to keep going." Four books refer to the motivation created by building assets even while paying down debt. Bach (2004, p. 204) writes, "If you were to direct all of your available cash flow to debt reduction . . . it might literally be years before you could begin saving for the future. This is too negative—so negative, in fact, that many people who follow this path get discouraged, give up early, and never get to the saving part." Three books endorse building up "long-term" savings in particular while paying down debt, ignoring the return differential between borrowed money and invested money. Lowry (2017, p. 215) writes, "Why bother saving when you have debt? Because trying to play catch-up later is a pain! Did that compound interest example show you nothing?!"

[9] Tyson (2019, p. 76) writes, "On the other hand, if you use savings to pay down credit-card debt, you can run your credit-card balances back up in a financial pinch (unless your card gets canceled)."

Mortgage Choices

Fixed versus Adjustable-Rate

Fixed-rate mortgages are exposed to inflation risk. Borrowers with a fixed-rate mortgage are better off under unexpectedly high inflation because it erodes the real present value of their loan repayments. Borrowers can protect themselves against a drop in inflation that lowers nominal interest rates by refinancing their fixed-rate mortgage, although many borrowers with a fixed-rate mortgage fail to refinance optimally (Keys, Pope, and Pope 2016).

In contrast, the real present value of adjustable-rate mortgage payments is almost unaffected by inflation, because changes in expected inflation change nominal interest rates in an adjustable-rate mortgage roughly one-for-one. However, borrowers with an adjustable-rate mortgage are exposed to the risk that real interest rates will change. They are also exposed to short-run variability in real mortgage payments; an increase in expected future inflation raises interest payments today even though the price level has not risen yet. Adjustable-rate mortgages usually charge lower interest rates than those with fixed rates, because interest rates of an adjustable-rate mortgage are pegged to short-term interest rates, whereas interest rates of a fixed-rate mortgage are pegged to long-term interest rates and include a premium for offering the refinancing option.

Weighing the above considerations, Campbell and Cocco (2003, 2015) find that borrowers should generally prefer adjustable-rate over fixed-rate mortgages, unless interest rates are low. Guren, Krishnamurthy, and McQuade (2021) find that adjustable-rate mortgages are also better than fixed-rate mortgages for macroeconomic stability, because short-term interest rates tend to fall more than long-term interest rates during recessions and fixed-rate mortgages require the borrower to refinance in order to obtain payment relief, which they will be unable to do if their home value has fallen enough to cause maximum loan-to-value requirements to bind.

Twenty-four books give advice about making choices about mortgages. The purported macroeconomic stabilization benefits of adjustable-rate mortgages notwithstanding, 11 books say that adjustable-rate mortgages are riskier than fixed-rate mortgages, with discussion focusing on the fluctuating monthly payments of adjustable-rate mortgages. The absence of discussion of inflation suggests that the perceived safety of fixed-rate mortgages is driven in part by money illusion; only two books mention that fixed-rate mortgages are exposed to inflation risk, but they see this exposure as advantageous—either as a hedge or a profit opportunity. Given views on the risks of adjustable-rate mortgages, it is not surprising that eight books recommend choosing fixed-rate instead of adjustable-rate mortgages. Only two books recommend choosing hybrid adjustable-rate mortgages which charge a fixed rate for a period of time before shifting to an adjustable rate, but they both advise avoiding exposure to the floating interest rate phase of the contract by choosing an initial fixed-rate period that corresponds to how long you plan to stay in the home.

Down Payment

Four books write approvingly of obtaining a mortgage with a 5 percent down payment or less in order to become a homeowner sooner, but all of these books were published before 2008. Five books recommend trying to make a down payment of at least 20 percent of the home's purchase price. None of the books in the sample suggest decreasing your down payment if you are pessimistic about housing returns in order to limit your financial exposure to house prices, as recommended by the risk-shifting model of Bailey et al. (2019) when homeowners are constrained from adjusting the size of their house in response to pessimism.[10]

Mortgage Term

Six books recommend taking a 30-year mortgage, citing the flexibility created by the lower monthly payments and the ability to lock in an interest rate for 30 years. Three books, all by Dave Ramsey, recommend a 15-year term. Ramsey (2013, p. 173) writes, "The really interesting thing I have observed is that fifteen-year mortgages always pay off in fifteen years . . . Thirty-year mortgages are for people who enjoy slavery so much they want to extend it for fifteen more years and pay thousands of dollars more for the privilege."

Mortgage Prepayment and Refinancing

Paying off your mortgage ahead of schedule is recommended by 11 books. Although the interest savings from doing so is mentioned by seven books, four books cite the emotional reward from owning your house debt-free as a reason. On the other hand, one book recommends against accelerating mortgage payments, citing higher expected returns from investing in the stock market, and five books are ambivalent about whether you should repay more quickly. The academic literature offers little guidance on this question.

Advice on when to refinance a fixed-rate mortgage is found in only five books. Chilton (1998) recommends refinancing if interest rates fall by at least 1 percentage point. Tyson (2019, p. 303) writes that refinancing is optimal "if you can recover the expenses of the refinance within a few years" or if you will keep the property and mortgage for at least as long as it will take to recover the refinancing expenses. Olen and Pollack (2016) say that refinancing is rarely worthwhile if the interest rate has dropped by less than 1 percentage point and otherwise depends on your tax rate, the outstanding mortgage balance, and when you expect to move homes; they refer readers to consult calculators on the internet. Ramsey (2013, p. 173) writes that "the best time to refinance is when you can save on interest," while Roth (2010, p. 216) says that the "standard advice" to wait until interest rates have dropped 2 percentage points is obsolete because closing costs are lower now.

[10] Bailey et al. (2019) provide numerous examples of popular financial advice to follow such a risk-shifting strategy. The fact that it does not appear in my sample suggests that such advice has had limited penetration.

Popular advice is considerably less nuanced than the approximately optimal refinancing rule derived by Agarwal, Driscoll, and Laibson (2013). The optimal strategy is complicated because of the option value of waiting for the interest rate to potentially fall further before paying the refinancing cost. The interest rate threshold for refinancing depends on the standard deviation of the mortgage interest rate, the cost of refinancing, the discount rate for future cashflows, the outstanding mortgage balance, the marginal tax rate against which mortgage interest can be deducted, and the expected time until the borrower will sell the home.

Conclusion

Popular financial advice can deviate from normative economic theory because of fallacies. But popular financial advice has two strengths relative to economic theory. First, the recommended action is often easily computable by ordinary individuals; there is no need to solve a complex dynamic programming problem. Second, the advice takes into account difficulties individuals have in executing a financial plan due to, say, limited motivation or emotional reactions to circumstances. Therefore, popular advice may be more practically useful to the ordinary individual. Developing normative economic models with these features, rather than ceding this territory to non-economists, may be a fruitful direction for future research.

■ *I thank John Campbell, Joao Cocco, Erik Hurst, Neale Mahoney, Nina Pavcnik, Timothy Taylor, Heidi Williams, and seminar audiences at the Bank of Italy, Caltech, Cheung Kong Graduate School of Business, the Personal Finance Symposium at the University of Alabama at Birmingham, University of Leeds, and Yale for helpful comments. Rohan Angadi, Rob Brinkmann, and Vod Vilfort provided excellent research assistance through the Yale Herb Scarf Summer Research Opportunities in Economics program.*

Appendix Table 1
Personal Finance Books Included in the Sample

	Goodreads rank (as of May 2019)
Bach, David. 2002. *Smart Couples Finish Rich*. New York: Broadway Books.	36
Bach, David. 2002. *Smart Women Finish Rich*. 2nd ed. New York: Broadway Books.	29
Bach, David. 2004. *The Automatic Millionaire*. New York: Broadway Books.	9
Bernstein, William. 2010. *The Four Pillars of Investing*. New York: McGraw-Hill.	14
Bernstein, William. 2017. *The Intelligent Asset Allocator*. 2nd ed. New York: McGraw-Hill.	47
Bogle, John. 1999. *Common Sense on Mutual Funds*. New York: John Wiley & Sons.	38
Bogle, John. 2017. *The Little Book of Common Sense Investing*. 10th anniversary ed. Hoboken: John Wiley & Sons, Inc.	11
Chilton, David. 1998. *The Wealthy Barber*. Updated 3rd edition. Roseville: Prima Publishing.	19
Chilton, David. 2011. *The Wealthy Barber Returns*. Kitchener: Financial Awareness Corp.	34
Clason, George. 1926. *The Richest Man in Babylon*. New York: New American Library, 1988.	6
Collins, J.L. 2016. *The Simple Path to Wealth*. Scotts Valley: CreateSpace Independent Publishing Platform.	15
Dacyczyn, Amy. 1998. *The Complete Tightwad Gazette*. New York: Villard Books.	48
DeMarco, M.J. 2018. *The Millionaire Fastlane*. Fountain Hills: Viperion Publishing Corporation.	39
Eker, T. Harv. 2005. *Secrets of the Millionaire Mind*. New York: HarperBusiness.	21
Ferri, Richard. 2010. *All About Asset Allocation*. 2nd ed. New York: McGraw-Hill.	46
Fisker, Jacob Lund. 2010. *Early Retirement Extreme*. Scotts Valley: CreateSpace Independent Publishing Platform.	26
Graham, Benjamin, and Jason Zweig. 2003. *The Intelligent Investor*. 4th revised ed., updated with new commentary by Jason Zweig. New York: HarperBusiness.	7
Hallam, Andrew. 2017. *Millionaire Teacher*. 2nd ed. Hoboken: John Wiley & Sons.	27
Kiyosaki, Robert T. 2017. *Rich Dad Poor Dad*. 20th anniversary ed. Scottsdale: Plata Publishing.	2
Kiyosaki, Robert T. 2012. *Rich Dad's Cashflow Quadrant*. 1st Plata Publishing edition. Scottsdale: Plata Publishing.	16
Kobliner, Beth. 2017. *Get A Financial Life*. New York: Simon & Schuster.	20
Lindauer, Mel, Taylor Larimore, and Michael LeBoeuf. 2014. *The Bogleheads' Guide to Investing*. 2nd ed. Hoboken: John Wiley & Sons.	12
Lowry, Erin. 2017. *Broke Millennial*. New York: Penguin Random House.	33
Lynch, Peter. 1989. *One Up on Wall Street*. New York: Simon & Schuster.	32
Malkiel, Burton G. 2019. *A Random Walk Down Wall Street*. New York: W. W. Norton & Company.	10
Mecham, Jesse. 2017. *You Need a Budget*. New York: HarperBusiness.	28
Olen, Helaine, and Harold Pollack. 2016. *The Index Card*. New York: Portfolio/Penguin.	23

(continued)

Appendix Table 1 (continued)
Personal Finance Books Included in the Sample

	Goodreads rank (as of May 2019)
Orman, Suze. 2007a. *The Money Book for the Young, Fabulous & Broke*. New York: Riverhead Books.	35
Orman, Suze. 2007b. *Women & Money*. New York: Spiegel & Grau.	13
Orman, Suze. 2012. *The 9 Steps to Financial Freedom*. 3rd paperback ed. New York: Currency.	37
Ramsey, Dave. 2003. *Financial Peace Revisited*. New York: Viking.	25
Ramsey, Dave. 2011. *Dave Ramsey's Complete Guide to Money*. Brentwood: Ramsey Press.	43
Ramsey, Dave. 2013. *The Total Money Makeover*, classic edition. Nashville: Nelson Books.	3
Richards, Carl. 2015. *The One-Page Financial Plan*. New York: Portfolio/Penguin.	50
Robbins, Tony. 2014. *Money: Master the Game*. New York: Simon & Schuster.	18
Robbins, Tony. 2017. *Unshakeable*. New York: Simon & Schuster Paperbacks.	22
Robin, Vicki, and Joe Dominguez. 2018. *Your Money or Your Life*. New York: Penguin Books.	4
Roth, J.D. 2010. *Your Money: The Missing Manual*. Sebastopol: O'Reilly Media.	49
Sethi, Ramit. 2019. *I Will Teach You to Be Rich*. 2nd ed. New York: Workman Publishing.	5
Sincero, Jen. 2017. *You Are a Badass at Making Money*. New York: Viking.	41
Stanley, Thomas J. 2001. *The Millionaire Mind*. Kansas City: Andrew McMeel Publishing	17
Stanley, Thomas J. 2009. *Stop Acting Rich... And Start Living Like a Real Millionaire*. Hoboken: John Wiley & Sons.	42
Stanley, Thomas, and William Danko. 1996. *The Millionaire Next Door*. Lanham: Taylor Trade Publishing.	1
Thames, Elizabeth Willard. 2018. *Meet the Frugalwoods*. New York: HarperBusiness.	44
Tobias, Andrew. 2016. *The Only Investment Guide You'll Ever Need*. 2nd Mariner Books ed. New York: Houghton Mifflin Harcourt Publishing Company.	24
Tyson, Eric. 2019. *Personal Finance for Dummies*. 9th ed. Hoboken: John Wiley & Sons.	31
Warren, Elizabeth, and Amelia Warren Tyagi. 2005. *All Your Worth*. New York: Free Press.	30
Books that contain no advice on covered topics	
Ferris, Timothy. 2009. *The 4-Hour Workweek*. Expanded and updated ed. New York: Crown Publishers.	40
Hill, Napoleon. 1937. *Think and Grow Rich*. Holden-Crowther Publications, 2018.	8
Wattles, Wallace. 1910. *The Science of Getting Rich*. New York: Penguin Group, 2007.	45

References

Agarwal, Sumit, John C. Driscoll, and David I. Laibson. 2013. "Optimal Mortgage Refinancing: A Closed-Form Solution." *Journal of Money, Credit and Banking* 45 (4): 591–622.

Aguiar, Mark, and Erik Hurst. 2005. "Consumption versus Expenditure." *Journal of Political Economy* 113 (5): 919–48.

Ameriks, John, and Stephen P. Zeldes. 2004. "How Do Household Portfolio Shares Vary with Age?" Unpublished.

Angeletos, George-Marios, David Laibson, Andrea Repetto, Jeremy Tobacman, and Stephen Weinberg. 2001. "The Hyperbolic Consumption Model: Calibration, Simulation, and Empirical Evaluation." *Journal of Economic Perspectives* 15 (3): 47–68.

Bach, David. 2004. *The Automatic Millionaire: A Powerful One-Step Plan to Live and Finish Rich.* New York: Broadway Books.

Bailey, Michael, Eduardo Dávila, Theresa Kuchler, and Johannes Stroebel. 2019. "House Price Beliefs and Mortgage Leverage Choice." *Review of Economic Studies* 86 (6): 2403–52.

Baker, H. Kent, and Rob Weigand. 2015. "Corporate Dividend Policy Revisited." *Managerial Finance* 41 (2): 126–44.

Barberis, Nicholas. 2000. "Investing for the Long Run When Returns Are Predictable." *Journal of Finance* 55 (1): 225–64.

Bekaert, Geert, Robert J. Hodrick, and Xiaoyan Zhang. 2009. "International Stock Return Comovements." *Journal of Finance* 64 (6): 2591–626.

Bender, Svetlana, James J. Choi, Danielle Dyson, and Adriana Z. Robertson. Forthcoming. "Millionaires Speak: What Drives Their Personal Investment Decisions?" *Journal of Financial Economics***.**

Bernheim, B. Douglas, Jonathan Skinner, and Steven Weinberg. 2001. "What Accounts for the Variation in Retirement Wealth among U.S. Households?" *American Economic Review* 91 (4): 832–57.

Bernstein, William. 2010. *The Four Pillars of Investing: Lessons for Building a Winning Portfolio.* New York: McGraw-Hill.

Bernstein, William. 2017. *The Intelligent Asset Allocator: How to Build Your Portfolio to Maximize Returns and Minimize Risk.* 2nd ed. New York: McGraw-Hill.

Bertaut, Carol C., Michael Haliassos, and Michael Reiter. 2009. "Credit Card Debt Puzzles and Debt Revolvers for Self Control." *Review of Finance* 13 (4): 657–92.

Biggs, Andrew G., and Glen R. Springstead. 2008. "Alternate Measures of Replacement Rates for Social Security Benefits and Retirement Income." *Social Security Bulletin* 68 (2): 1–19.

Board of Governors of the Federal Reserve System (US). Commercial Bank Interest Rate on Credit Card Plans, Accounts Assessed Interest [TERMCBCCINTNS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/TERMCBCCINTNS, September 22, 2022.

Bodie, Zvi, Robert C. Merton, and William F. Samuelson. 1992. "Labor Supply Flexibility and Portfolio Choice in a Life Cycle Model." *Journal of Economic Dynamics and Control* 16 (3-4): 427–49.

Bogle, John. 1999. *Common Sense on Mutual Funds: New Imperatives for the Intelligent Investor.* New York: John Wiley & Sons.

Bogle, John. 2017. *The Little Book of Common Sense Investing: The Only Way to Guarantee Your Fair Share of Stock Market Returns.* 10th ed. Hoboken: John Wiley & Sons.

Brennan, Michael J., and Yihong Xia. 2000. "Stochastic Interest Rates and the Bond-Stock Mix." *European Finance Review* 4 (2): 197–210.

Campbell, John Y., and João F. Cocco. 2003. "Household Risk Management and Optimal Mortgage Choice." *Quarterly Journal of Economics* 118 (4): 1449–94.

Campbell, John Y., and João F. Cocco. 2015. "A Model of Mortgage Default." *Journal of Finance* 70 (4): 1495–554.

Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay. 1997. *The Econometrics of Financial Markets.* Princeton: Princeton University Press.

Campbell, John Y., and Roman Sigalov. 2022. "Portfolio Choice with Sustainable Spending: A Model of Reaching for Yield." *Journal of Financial Economics* 143 (1): 188–206.

Campbell, John Y., and Luis M. Viceira. 2001. "Who Should Buy Long-Term Bonds?" *American Economic Review* 91 (1): 99–127.

Canner, Niko, N. Gregory Mankiw, and David N. Weil. 1997. "An Asset Allocation Puzzle." *American Economic Review* 87 (1): 181–91.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52 (1): 57–82.
Carroll, Christopher D. 1997. "Buffer-Stock Saving and the Life Cycle/Permanent Income Hypothesis." *Quarterly Journal of Economics* 112 (1): 1–55.
Chevalier, Judith, and Glenn Ellison. 1997. "Risk Taking by Mutual Funds as a Response to Incentives." *Journal of Political Economy* 105 (6): 1167–200.
Chilton, David. 1998. *The Wealthy Barber: Everyone's Commonsense Guide to Becoming Financially Independent*. 3rd ed. Roseville: Prima Publishing.
Chilton, David. 2011. *The Wealthy Barber Returns: Significantly Older and Marginally Wiser, Dave Chilton Offers His Unique Perspectives on the World of Money*. Kitchener: Financial Awareness Corp.
Choi, James J., and Adriana Z. Robertson. 2020. "What Matters to Individual Investors? Evidence from the Horse's Mouth." *Journal of Finance* 75 (4): 1965–2020.
Choi, James J., and Kevin Zhao. 2021. "Carhart (1997) Mutual Fund Performance Persistence Disappears Out of Sample." *Critical Finance Review* 10 (2): 263–70.
Chopra, Felix. 2021. "Media Persuasion and Consumption: Evidence from the Dave Ramsey Show." Unpublished.
Christoffersen, Peter, Vihang Errunza, Kris Jacobs, and Hugues Langlois. 2012. "Is the Potential for International Diversification Disappearing? A Dynamic Copula Approach." *Review of Financial Studies* 25 (12): 3711–51.
Clason, George S. 1926. *The Richest Man in Babylon*. New York: New American Library, 1988.
Cocco, João F., Francisco J. Gomes, and Pascal J. Maenhout. 2005. "Consumption and Portfolio Choice over the Life Cycle." *Review of Financial Studies* 18 (2): 491–533.
Cochrane, John H. 2009. *Asset Pricing: Revised Edition*. Princeton: Princeton University Press.
Coeurdacier, Nicolas, and Hélène Rey. 2013. "Home Bias in Open Economy Financial Macroeconomics." *Journal of Economic Literature* 51 (1): 63–115.
Collins, J. L. 2016. *The Simple Path to Wealth: Your Road Map to Financial Independence and a Rich, Free Life*. Scotts Valley: CreateSpace Independent Publishing Platform.
Dacyczyn, Amy. 1998. *The Complete Tightwad Gazette: Promoting Thrift as a Viable Alternative Lifestyle*. New York: Villard Books.
Fama, Eugene F., and Kenneth R. French. 1992. "The Cross-Section of Expected Stock Returns." *Journal of Finance* 47 (2): 427–65.
Fama, Eugene F., and Kenneth R. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.
Federal Deposit Insurance Corporation. National Deposit Rates: Savings [SNDR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SNDR, September 22, 2022.
Ferri, Richard A. 2010. *All About Asset Allocation*. 2nd ed. New York: McGraw-Hill.
French, Kenneth R. 2008. "Presidential Address: The Cost of Active Investing." *Journal of Finance* 63 (4): 1537–73.
French, Kenneth R. 2022a. "Portfolios Formed on Dividend Yield." https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/Portfolios_Formed_on_D-P_TXT.zip (accessed August 3, 2022).
French, Kenneth R. 2022b. "FamaFrench 3 Factors." https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/ftp/F-F_Research_Data_Factors_TXT.zip (accessed August 3, 2022).
French, Kenneth R., and James M. Poterba. 1991. "Investor Diversification and International Equity Markets." *American Economic Review* 81 (2): 222–6.
Fulford, Scott L. 2015. "How Important is Variability in Consumer Credit Limits?" *Journal of Monetary Economics* 72 (C): 42–63.
Gathergood, John, Neale Mahoney, Neil Stewart, and Jörg Weber. 2019. "How Do Individuals Repay Their Debt? The Balance-Matching Heuristic." *American Economic Review* 109 (3): 844–75.
Gomes, Francisco, and Alexander Michaelides. 2005. "Optimal Life-Cycle Asset Allocation: Understanding the Empirical Evidence." *Journal of Finance* 60 (2): 869–904.
Gourinchas, Pierre-Olivier, and Jonathan A. Parker. 2002. "Consumption over the Life Cycle." *Econometrica* 70 (1): 47–89.
Gross, David B., and Nicholas S. Souleles. 2002. "Do Liquidity Constraints and Interest Rates Matter for Consumer Behavior? Evidence from Credit Card Data." *Quarterly Journal of Economics* 117 (1): 149–85.
Guiso, Luigi, and Paolo Sodini. 2013. "Household Finance: An Emerging Field." In *Handbook of the Economics of Finance,* edited by George Constantinides, Milton Harris, and Rene M. Stulz, 1397–532. Amsterdam: Elsevier.

Guren, Adam M., Arvind Krishnamurthy, and Timothy J. McQuade. 2021. "Mortgage Design in an Equilibrium Model of the Housing Market." *Journal of Finance* 76 (1): 113–68.

Haliassos, Michael, and Carol C. Bertaut. 1995. "Why Do So Few Hold Stocks?" *Economic Journal* 105 (432): 1110–29.

Hartzmark, Samuel M., and David H. Solomon. 2019. "The Dividend Disconnect." *Journal of Finance* 74 (5): 2153–99.

Investment Company Institute. 2021. *Investment Company Fact Book: A Review of Trends and Activities in the Investment Company Industry*. Washington, DC: Investment Company Institute.

Kaplan, Greg, and Giovanni L. Violante. 2014. "A Model of Consumption Response to Fiscal Stimulus Payments." *Econometrica* 82 (4): 1199–239.

Kaplan, Greg, Giovanni L. Violante, and Justin Weidner. 2014. "The Wealthy Hand-to-Mouth." *Brookings Papers on Economic Activity* 45 (1): 77–138.

Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman. 2016. "Getting to the Top of Mind: How Reminders Increase Saving." *Management Science* 62 (12): 3393–411.

Keys, Benjamin J., Devin G. Pope, and Jaren C. Pope. 2016. "Failure to Refinance." *Journal of Financial Economics* 122 (3): 482–99.

Kiyosaki, Robert T. 2012. *Rich Dad's Cashflow Quadrant: Guide to Financial Freedom*. Scottsdale: Plata Publishing.

Kiyosaki, Robert T. 2017. *Rich Dad Poor Dad: What the Rich Teach Their Kids about Money that the Poor and Middle Class Do Not!* Scottsdale: Plata Publishing.

Kobliner, Beth. 2017. *Get A Financial Life: Personal Finance in Your Twenties and Thirties*. New York: Simon & Schuster.

Lehnert, Andreas, and Dean Maki. 2007. "Consumption, Debt and Portfolio Choice: Testing the Effect of Bankruptcy Law." In *Household Credit Usage*, edited by Sumit Agarwal and Brent W. Ambrose, 55–76. New York: Palgrave Macmillan.

Lewis, Karen K. 1999. "Trying to Explain Home Bias in Equities and Consumption." *Journal of Economic Literature* 37 (2): 571–608.

Lisa, Andrew. 2021. "12 of the Bestselling Financial Advice Books of All Time." *GOBankingRates*, April 23. https://www.gobankingrates.com/money/financial-planning/12-of-the-bestselling-financial-advice-books-of-all-time/.

Lynch, Peter. 1989. *One Up on Wall Street: How to Use What You Already Know to Make Money in the Market*. New York: Simon & Schuster.

Lowry, Erin. 2017. *Broke Millennial: Stop Scraping By and Get Your Financial Life Together*. New York: Penguin Random House.

Malkiel, Burton G. 2019. *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*. New York: W. W. Norton & Company.

Mehra, Rajnish, and Edward C. Prescott. 1985. "The Equity Premium: A Puzzle." *Journal of Monetary Economics* 15 (2): 145–61.

Merton, Robert C. 1969. "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case." *Review of Economics and Statistics* 51 (3): 247–57.

Miller, Merton H., and Franco Modigliani. 1961. "Dividend Policy, Growth, and the Valuation of Shares." *Journal of Business* 34 (4): 411–33.

NPD. 2020. "Personal Finance Book Sales Are on the Rise, the NPD Group Says." NPD, September 21. https://www.npd.com/news/press-releases/2020/personal-finance-book-sales-are-on-the-rise-the-npd-group-says/ (Accessed August 1, 2022).

Olen, Helaine, and Harold Pollack. 2016. *The Index Card: Why Personal Finance Doesn't Have to Be Complicated*. New York: Portfolio/Penguin.

Orman, Suze. 2007a. *The Money Book for the Young, Fabulous & Broke*. New York: Riverhead Books.

Orman, Suze. 2007b. *Women & Money*. New York: Spiegel & Grau.

Orman, Suze. 2012. *The 9 Steps to Financial Freedom: Practical & Spiritual Steps So You Can Stop Worrying*. 3rd ed. New York: Currency.

Pástor, Luboš, and Robert F. Stambaugh. 2012. "Are Stocks Really Less Volatile in the Long Run?" *Journal of Finance* 67 (2): 431–77.

Perold, André F., and Evan C. Schulman. 1988. "The Free Lunch in Currency Hedging: Implications for Investment Policy and Performance Standards." *Financial Analysts Journal* 44 (3): 45–50.

Rabin, Matthew. 2000. "Risk Aversion and Expected-Utility Theory: A Calibration Theorem." *Econometrica* 68 (5): 1281–92.

Ramsey, Dave. 2003. *Financial Peace Revisited.* New York: Viking.
Ramsey, Dave. 2011. *Dave Ramsey's Complete Guide to Money: The Handbook of Financial Peace University.* Brentwood: Ramsey Press.
Ramsey, Dave. 2013. *The Total Money Makeover: A Proven Plan for Financial Fitness.* Nashville: Nelson Books.
Robbins, Tony. 2014. *Money—Master the Game: 7 Simple Steps to Financial Freedom.* New York: Simon & Schuster.
Robin, Vicki, and Joe Dominguez. 2018. *Your Money or Your Life: 9 Steps to Transforming Your Relationships with Money and Achieving Financial Independence.* New York: Penguin Books.
Roth, J.D. 2010. *Your Money: The Missing Manual.* Sebastopol: O'Reilly Media.
Samuelson, Paul A. 1969. "Lifetime Portfolio Selection by Dynamic Stochastic Programming." *Review of Economics and Statistics* 51 (3): 239–46.
Securities Industry and Financial Markets Association (SIFMA). 2021. *SIFMA Research Quarterly—3Q21.* New York: Securities Industry and Financial Markets Association.
Telyukova, Irina A., and Randall Wright. 2008. "A Model of Money and Credit, with Application to the Credit Card Debt Puzzle." *Review of Economic Studies* 75 (2): 629–47.
Thaler, Richard H. 1985. "Mental Accounting and Consumer Choice." *Marketing Science* 4 (3): 199–214.
Thaler, Richard H. 1999. "Mental Accounting Matters." *Journal of Behavioral Decision Making* 12 (3): 183–206.
Tyson, Eric. 2019. *Personal Finance for Dummies.* 9th ed. Hoboken: John Wiley & Sons.
Vissing-Jørgensen, Annette. 2003. "Perspectives on Behavioral Finance: Does 'Irrationality' Disappear with Wealth? Evidence from Expectations and Actions." *NBER Macroeconomics Annual* 18: 139–94.
Wachter, Jessica A. 2002. "Portfolio and Consumption Decisions under Mean-Reverting Returns: An Exact Solution for Complete Markets." *Journal of Financial and Quantitative Analysis* 37 (1): 63–91.
Wachter, Jessica A. 2003. "Risk Aversion and Allocation to Long-Term Bonds." *Journal of Economic Theory* 112 (2): 325–33.
Warren, Elizabeth, and Amelia Warren Tyagi. 2005. *All Your Worth: The Ultimate Lifetime Money Plan.* New York: Free Press.
Yaari, Menahem E. 1965. "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer." *Review of Economic Studies* 32 (2): 137–50.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 193–204

A Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure

Liyang Sun and Jesse M. Shapiro

Economists often seek to evaluate the effects of a certain event, such as the adoption of a policy or the arrival of an innovation, on some outcome of interest. For example, how did the enactment of Medicare (government health insurance for all elderly Americans) affect total expenditures on health care? How did the historical arrival of the potato affect population growth across the Old World? Simply comparing outcomes before and after the occurrence of the event risks conflating the effect of the event with the effect of numerous other coincident changes: think of all the other things that changed around the start of Medicare (1965) or the beginning of the Columbian exchange (1492). One way to measure the effect of these coincident changes is by looking at the outcomes of a control group totally unaffected by the event. But in some cases it is difficult to find such a pure control—Medicare was a national policy, and the arrival of the potato likely touched every part of the Old World in some way.

In such settings, it is common to take advantage of variation across geographic or other units in the extent of their *exposure* to the event. Even though all US states were affected by the introduction of Medicare, some were more affected than others, for example, because they had relatively less well-insured elderly populations prior

■ *Liyang Sun is Postdoctoral Research Fellow, University of California Berkeley, Berkeley, California and Assistant Professor, CEMFI (Center for Monetary and Financial Studies), Madrid, Spain. Jesse M. Shapiro is George Gund Professor of Economics and Business Administration, Harvard University, Cambridge, Massachusetts and Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are lsun20@berkeley.edu and jesse_shapiro@fas.harvard.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.193.

to Medicare. Likewise, some regions of the Old World were relatively better suited to potato cultivation, making them better able to take advantage of the new crop's arrival.

One model of such a situation holds that the outcome is composed of a unit effect, a time effect, an interaction between a measure of the event and a measure of the unit's exposure, and an error term unrelated to the others. We can write a heuristic model like this:

$$\text{Outcome} = \text{Unit effect} + \text{Time effect} + \text{Coefficient}\ (\text{Event} \times \text{Exposure}) + \text{Error}. \quad \text{(heuristic model)}$$

In this linear panel model, the unknown unit effect accounts for features of the unit (for example, state or region) that are time-invariant, the unknown time effect accounts for background changes that may coincide with the event, and the unknown error term accounts for other unsystematic factors that influence different units at different times. The observed event variable varies over time and captures the event of interest. The observed exposure variable varies across units and captures units' different exposure to the event. The product of these two variables is the term of greatest interest, as it captures the fact that different units are affected differently by the event because of their different exposure to it.

Linear panel models featuring an interaction between an event variable and an exposure variable, as in the heuristic model, appear in many areas of economics. For example, Finkelstein's (2007, equation 1) model of hospital expenses includes an interaction between time indicators (around the introduction of Medicare) and a measure of access to private insurance. Nunn and Qian's (2011, equation 3) model of Old World population growth includes an interaction between an indicator for periods following the introduction of the potato and the log of land area in a country that is suitable for growing potatoes. Dube and Vargas' (2013, equation 1) model of violence in Colombia includes an interaction between the world oil price and a measure of a region's baseline oil production intensity.[1]

Under suitable conditions on the error term, the unknown coefficient in the heuristic model can be estimated via ordinary least squares regression of the outcome on unit indicators, time indicators, and an interaction between the event variable and the exposure variable. Because the model involves two sets of fixed effects—one for units and one for time—this ordinary least squares estimator is sometimes called a two-way fixed effects (TWFE) estimator.

In this paper we consider the possibility that, in addition to the exposure variable, the effect of the policy or event itself—the coefficient in the heuristic model—differs by unit. Heterogeneous effects of this kind can arise for many reasons. For example,

[1] Other examples include Zhang and Zhu's (2011, equations 2 and 3) model of social influences on contributions to Chinese Wikipedia; Dafny, Duggan, and Ramanarayanan's (2012, equation 5) model of the effect of a merger on health insurance premiums; and Pierce and Schott's (2016, equation 2) model of the effect of trade with China on US manufacturing employment.

a given change in the fraction of elderly insured might affect expenditures more in states with a less healthy uninsured population. A given level of potato cultivation might affect population growth more in regions with better access to trade. Economists have been interested in heterogeneous effects of this kind for a long time (see, for example, surveys in Heckman and Vytlacil 2007; Imbens and Wooldridge 2009). Recently, an especially active literature has studied the effects of this form of coefficient heterogeneity on the performance and interpretation of the two-way fixed effects estimator. We draw heavily on this literature, and especially on work by de Chaisemartin and D'Haultfœuille (2018), who consider a setting similar to the one we consider here.

We will see that in general the two-way fixed effects estimator can perform very poorly when effects are heterogeneous, in the sense that it can fail to estimate the average (or even a weighted average) of the units' coefficients. This problem can be so severe that it affects any estimator, not just the TWFE estimator. And we will look at one situation—a setting with an unaffected unit—in which it is possible to estimate an average effect by replacing the TWFE with an average of exposure-adjusted difference-in-differences estimators.

A Motivating Example

To study the issues in more detail, we now introduce a concrete example. We base the example loosely on Finkelstein's study of the effect of Medicare, setting aside much of the richness of Finkelstein's (2007) original analysis.

We are interested in learning the effect of Medicare on health care expenditures. Medicare is a US government program introduced in 1965 to provide health insurance to the elderly. We observe per capita health care expenditures y_{st} on the elderly for each US state s in each of two time periods t, where we can let $t = 0$ denote the period before the introduction of Medicare and $t = 1$ denote the period after.

Because many things change over time, simply comparing expenditures at time $t = 1$ to those at time $t = 0$ may not give a reliable estimate of the effect of Medicare. It would be helpful to have a control state that did not adopt Medicare, but since Medicare was a national policy, such a state does not exist.

Instead, we can take advantage of the fact that states differ in the fraction of the elderly that were insured prior to Medicare's introduction. In a New England state, where the penetration of private insurance among the elderly was relatively high prior to the introduction of Medicare (Finkelstein 2007, Table 1), Medicare had a relatively small effect on rates of insurance coverage. In a Pacific state, where the penetration of private insurance among the elderly was relatively low prior to Medicare (Finkelstein 2007, Table 1), Medicare had a relatively large effect on rates of insurance coverage.

Let x_{st} be the fraction of elderly with health insurance in a given state s at time t. At time $t = 0$, before Medicare, we can think of x_{s0} as measuring the fraction of

elderly with private or other (non-Medicare) government insurance in state *s*. At time $t = 1$, after Medicare, we can think of x_{s1} as being equal to 1 for all states *s* due to the universal coverage afforded by Medicare.

A linear panel data model of health care expenditures—what we will refer to as the linear model—might then take the form

$$y_{st} = \alpha_s + \delta_t + \beta x_{st} + \varepsilon_{st}. \qquad \text{(linear model)}$$

Here α_s is a state fixed effect that captures time-invariant state characteristics that may affect health care expenditures, δ_t is a time fixed effect that captures state-invariant time-dependent factors that may affect health care expenditures, and ε_{st} is an error term unrelated to x_{st}.[2] The parameter β measures the causal effect of insurance coverage on health care expenditure. Specifically, it measures the effect on per capita health expenditures of going from no coverage ($x_{st} = 0$) to full coverage ($x_{st} = 1$).[3]

We can rewrite the linear model in a form that resembles the heuristic model. In particular, because $x_{s1} = 1$ for all states *s*, it is straightforward to show that the linear model implies that

$$y_{st} = \tilde{\alpha}_s + \delta_t + \beta(1 - x_{s0})\, t + \varepsilon_{st}. \qquad \text{(exposure model)}$$

In the exposure model, the term $\tilde{\alpha}_s$ plays the role of the unknown unit effect from the heuristic model.[4] The term δ_t plays the role of the unknown time effect. The term ε_{st} plays the role of the unknown error term. The term $(1 - x_{s0})$ is the observed exposure variable and the term *t*, which is just an indicator for whether the observation is from the post-Medicare period, is the observed event variable.

Intuitively, under the exposure model, we can learn about the coefficient β by looking at whether, following the introduction of Medicare, health care expenditures diverge between states with different levels of private insurance before Medicare (different values of x_{s0}). If so, then because different states are affected equally by the time effect represented by δ_t, it must be that Medicare is exerting a causal effect on expenditures.

More practically, we can estimate the unknown coefficient β by regressing health expenditures on state indicators, a time indicator, and an interaction between the fraction previously uninsured $(1 - x_{s0})$ and the post-Medicare indicator *t*. This is a two-way fixed effects estimator. Call it $\hat{\beta}$. The TWFE estimator $\hat{\beta}$ has some appealing properties. For example, if the exposure model holds, and ε_{st} is unrelated to x_{st}, then

[2] Specifically, we assume that each of ε_{s0} and ε_{s1} has mean zero conditional on x_{s0} and x_{s1}.

[3] The effect of Medicare on expenditures in state *s* is given by $\beta(1 - x_{s0})$, that is, the effect of insurance coverage on expenditures, β, multiplied by the effect of Medicare on insurance coverage, $(1 - x_{s0})$.

[4] To go from the linear model to the exposure model, we have redefined the state fixed effect as $\tilde{\alpha}_s = \alpha_s + \beta x_{s0}$.

$\hat{\beta}$ is centered around β, in the sense that even though in any given sample $\hat{\beta}$ may be higher or lower than β, across samples $\hat{\beta}$ will tend to be equal to β on average.

The Possibility of Heterogeneous Coefficients

According to the linear model, a given change in the fraction insured has the same effect on per capita health expenditures in every state *s*. But it seems plausible that health expenditures will respond differently to changes in insurance in different states. For example, a state with a less healthy uninsured population may see expenditures rise more in response to a given expansion in insurance, compared to a state with a more healthy uninsured population, because relatively less healthy insurees require more expensive care.

We can formalize this possibility by imagining that each state *s* has its own coefficient β_s describing the effect of insurance on expenditures in the state, much as it has its own fixed effect α_s describing its baseline level of expenditures. Keeping all other elements of the linear model yields the following heterogeneous panel model:

$$y_{st} = \alpha_s + \delta_t + \beta_s x_{st} + \varepsilon_{st}. \qquad \text{(heterogeneous model)}$$

Even though we are allowing heterogeneity in the effect of treatment, we are still maintaining that the error term ε_{st} is unrelated to the fraction of elderly with health insurance x_{st} as before, so absent changes in the insurance levels x_{st}, all states would follow identical average trends over time.

Consider a researcher who believes that the effect of insurance may differ across states as in the heterogeneous model. How reasonable would it be for the researcher to estimate the effect of added health insurance using the convenient two-way fixed effects estimator that is based on the exposure model, which assumes that all states have the same coefficient β?

A single estimator $\hat{\beta}$, by construction, cannot be centered around each of the 50 different true coefficients for each state β_s. But maybe the single estimator $\hat{\beta}$ is centered around a good summary of the true coefficients, such as an average. If so, $\hat{\beta}$ might still be a convenient way to estimate the effect of insurance on expenditures in a "typical" state.

In certain situations, the estimator $\hat{\beta}$ will indeed be centered on an average of the true state-level coefficients β_s. One such situation is where β_s is unrelated to (that is, statistically independent of) all the other terms in the heterogeneous model. In this case, results in the online Appendix imply that $\hat{\beta}$ is centered around an average of the coefficients β_s, and therefore might still be considered an appealing estimator.

However, the situation where the coefficient β_s is unrelated to the other terms in the model is somewhat special. Suppose, for example, that β_s is greater in states with a less healthy uninsured population. Suppose, further, that the uninsured

population is less healthy in states with greater insurance penetration prior to Medicare, say because in such states only the least healthy elderly remain uninsured. In this case, β_s will tend to be positively related to x_{s0}. Such a relationship between β_s and x_{s0} can cause the two-way fixed effects estimator $\hat{\beta}$ to behave rather badly.

To illustrate, consider a hypothetical numerical example of the heterogeneous model. In this example, we let the index s of the states run from 1 to 50. We let the fraction of elderly with insurance before Medicare be given by $x_{s0} = 0.245 + s/100$, so that the fraction runs from 0.255 ($s = 1$) through 0.745 ($s = 50$) in increments of 0.01, with an average value of 0.5.

In this numerical example, we also let the coefficient β_s vary across states according to the equation

$$\beta_s = 1 + 0.5\lambda - \lambda x_{s0}. \qquad \text{(numerical example)}$$

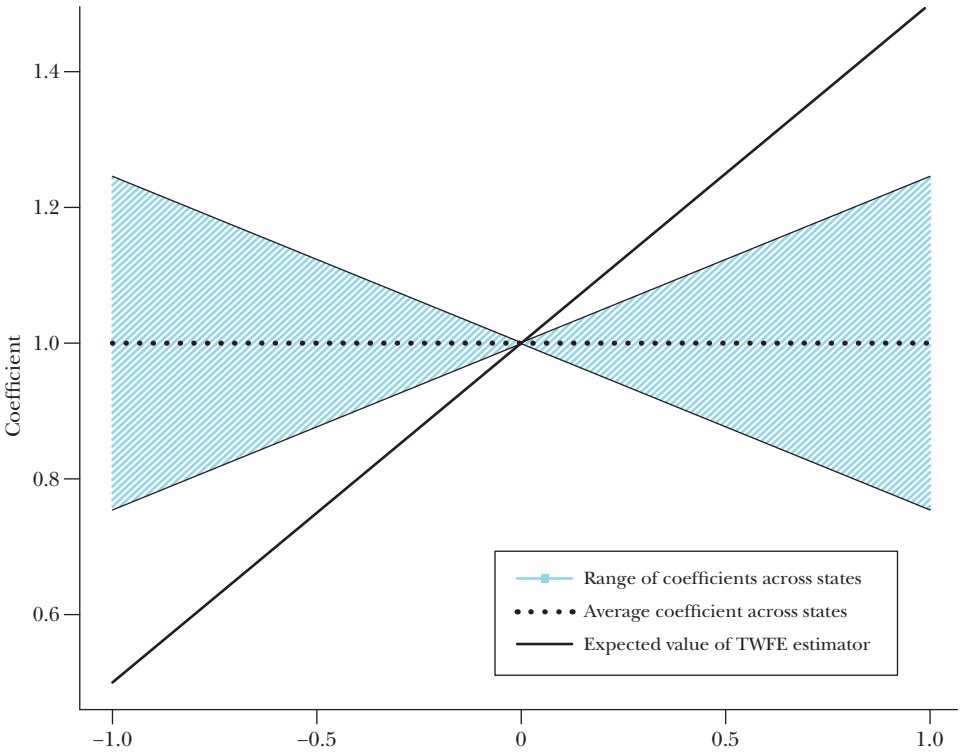
Here, λ is a parameter that governs how much the coefficient β_s varies across states, and how the state-level coefficient β_s is related to the fraction of elderly with insurance before Medicare. When λ is 0, the coefficient β_s is equal to 1 in all states regardless of prior insurance penetration. When λ is less than 0, states with greater insurance penetration prior to Medicare have a larger coefficient β_s. When λ is greater than 0, states with greater insurance penetration prior to Medicare have a smaller coefficient β_s.

We have constructed the numerical example so that, no matter the value of λ, the average value of β_s across all states is always 1. By varying λ, we can therefore vary the relationship between β_s and x_{s0} while holding constant the average value of β_s.

Figure 1 illustrates the behavior of the two-way fixed effects estimator $\hat{\beta}$ in this numerical example. The horizontal axis shows the parameter λ, which controls the strength of the relationship between β_s and x_{s0}, and hence the degree of heterogeneity in the coefficient β_s. We consider values of λ ranging from −1 to +1. The shaded region shows the range of coefficients across the states s for each given value of λ. As λ departs from zero, this range widens, but remains centered around the average value of 1, which is illustrated with a dotted line. The solid line shows the value around which the TWFE estimator $\hat{\beta}$ is centered. Specifically, the line shows the average or expected value of $\hat{\beta}$ across repeated samples of the data. Except when $\lambda = 0$, this value, which is derived in the online Appendix, does not coincide with the average value of β_s.

Perhaps more surprising, and more concerning, is that, when λ is not equal to zero, $\hat{\beta}$ is centered outside the shaded region that depicts the range of true coefficients β_s. When λ is less than zero, $\hat{\beta}$ is centered on a value smaller than any of the true coefficients β_s. When λ is greater than zero, $\hat{\beta}$ is centered on a value larger than any of the true coefficients β_s. A researcher using $\hat{\beta}$ to estimate an average or typical effect of insurance on health expenditure would, in these situations, end up with a very misleading estimate, one that is centered on a value outside the range of the true coefficients β_s.

Figure 1
Expected Value of the Two-Way Fixed Effects Estimator under Coefficient Heterogeneity



Source: Illustrative calculations by the authors.
Notes: This figure illustrates the behavior of the two-way fixed effects (TWFE) estimator $\hat{\beta}$ of the parameter β in the exposure model for a hypothetical numerical example described in Section "The Possibility of Heterogeneous Coefficients." The horizontal axis corresponds to the parameter λ which governs how much and in what way the coefficient β_s in the heterogeneous model varies across states. For each given value of λ, the shaded region shows the range of coefficients $(\beta_1, \ldots, \beta_{50})$ across the 50 states, the dotted line shows the average value of β_s, and the solid line shows expected value of the two-way fixed effects estimator, $\hat{\beta}$.

To understand why the two-way fixed effects estimator behaves this way, consider the case of $\lambda > 0$ and recall that $\hat{\beta}$ is the ordinary least squares estimate of the coefficient β on the interaction term $(1 - x_{s0})\,t$ in the exposure model. This estimate will tend to be larger when, following Medicare's introduction, expenditure grows more in states that experience a larger increase in insurance coverage, $(1 - x_{s0})$. When $\lambda > 0$, states with a larger increase in insurance coverage, $(1 - x_{s0})$, also have larger coefficients β_s. Following Medicare's introduction, expenditure therefore grows more in states with larger $(1 - x_{s0})$ both because these states experience a larger increase in insurance coverage and because these states experience a larger change in expenditure for a given change in insurance coverage. The exposure

model accounts for only the first of these effects, so the corresponding ordinary least squares estimator $\hat{\beta}$ conflates them, thus overstating the effect of insurance on expenditure. In the numerical example, this conflation is so severe that the expected value of the TWFE estimator falls outside the range of the true coefficients β_s.

The numerical example proves that the two-way fixed effects estimator cannot, in general, be guaranteed to be centered around a value inside the range of the true coefficients β_s in the heterogeneous model. In fact, we prove in the online Appendix that there is *no* estimator that can be guaranteed, regardless of the coefficients β_s and the pre-Medicare insurance levels x_{s0}, to be centered around a value inside the range of the true coefficients β_s in the heterogeneous model. It follows that there is no estimator guaranteed to be centered around the average β_s across the states. The proof focuses on the case where $x_{s0} < 1$ for all s, and so, as in the case of Medicare, there is no totally unaffected state.

A Difference-in-Differences Perspective

Another way to build intuition about the impact of coefficient heterogeneity is to consider the behavior of some difference-in-differences type estimators. To relate to the classical difference-in-differences estimator, imagine that Medicare had been adopted in one treatment state, say state s, and not adopted in another control state, say state s'. Imagine further that no one had health insurance to begin with in either state, so that Medicare increased the fraction of the elderly with health insurance from 0 to 1 in the treatment state s, and left the fraction at 0 in the control state s'. In this case, by computing the difference in the change in the outcome y between the treatment and control states, $(y_{s1} - y_{s0}) - (y_{s'1} - y_{s'0})$, we would, on average, isolate the effect of Medicare, and arrive at a difference-in-differences estimator centered around the true effect β, much as in Card and Krueger's (1994) classic study of the effect of the minimum wage.

In this hypothetical situation, we have one treatment state that is strongly affected by the introduction of Medicare, and another control state that is totally unaffected. In the more realistic situation where all states were affected by the introduction of Medicare, simply comparing the change in the outcome y between a more affected state s and a less affected state s' seems incomplete, because such a comparison does not account for the different changes in insurance rates x induced by Medicare in the two states. The following exposure-adjusted difference-in-differences estimator provides one possible way to account for changes in insurance rates:

$$\hat{\beta}^{DID}_{s,s'} = \frac{(y_{s1} - y_{s0}) - (y_{s'1} - y_{s'0})}{(1 - x_{s0}) - (1 - x_{s'0})}.$$

De Chaisemartin and D'Haultfœuille (2018) call $\hat{\beta}^{DID}_{s,s'}$ a Wald-difference-in-differences estimator because it consists of the ratio of the difference-in-differences estimator for the outcome (in our case, expenditures) to the one for exposure (insurance).

The estimator $\hat{\beta}^{DID}_{s,s'}$ is intuitive, but suffers from limitations similar to those of the two-way fixed effects estimator. In particular, $\hat{\beta}^{DID}_{s,s'}$ can be centered around a value that is larger or smaller than both β_s, the true coefficient for state s, and $\beta_{s'}$, the true coefficient for state s'. For a concrete example, if we take $s = 1$ and $s' = 50$ from the earlier numerical example, and say that $\lambda = 1$, then based on the formula we derived, the estimator $\hat{\beta}^{DID}_{s,s'}$ is centered around the value 1.5, which is greater than both $\beta_1 = 1.245$ and $\beta_{50} = 0.755$. One way to build an intuition for this behavior is to note that $\hat{\beta}^{DID}_{s,s'}$ is equivalent to the TWFE estimator $\hat{\beta}$ in the case where we have only two states in the sample, s and s'. Just like the TWFE estimator, $\hat{\beta}^{DID}_{s,s'}$ cannot be guaranteed to be centered around a value inside the range of β_s and $\beta_{s'}$.[5]

Suppose, though, that in state s' Medicare had no effect on insurance rates, for example, because all elderly in the state were insured prior to Medicare, $x_{s'0} = 1$. That would take us closer to the classical difference-in-differences setting of Card and Krueger (1994) and others, and in that case, $\hat{\beta}^{DID}_{s,s'}$ is centered around β_s, the true coefficient for the affected state s. In fact, by taking an average of $\hat{\beta}^{DID}_{s,s'}$ across all of the affected states s, always treating state s' as the comparison, we arrive at an estimator that is centered around the average value of β_s across all affected states s.[6]

The presence of a totally unaffected state therefore makes it possible to construct an estimator centered around the true coefficient for any affected state, such as $\hat{\beta}^{DID}_{s,s'}$, and one centered around the average of true coefficients for all affected states, such as the average of $\hat{\beta}^{DID}_{s,s'}$. It is important to note, however, that the presence of a totally unaffected state does not repair the problems we highlighted earlier with the two-way fixed effects estimator $\hat{\beta}$. Calculations in the online Appendix show that even if we add a totally unaffected state to the sample, the TWFE estimator remains centered outside of the range of treatment effects β_s in the numerical example. Thus, while the presence of a totally unaffected state means that it is possible to find estimators that are centered around the average coefficient, it does not guarantee that all estimators are centered around an average coefficient.

Some economic situations do not feature a totally unaffected unit that can serve as a comparison for affected units. In such situations, researchers may still be able to make progress by using economic assumptions to impose further structure on the coefficients β_s. For example, suppose that a researcher is willing to posit a linear relationship between β_s and x_{s0} of the form in the numerical example, but does not know the value of the parameter λ that governs this relationship. In this case, it is possible to substitute the expression for β_s into the heterogeneous model to arrive at a linear panel model whose unknown parameter, λ, can be estimated by

[5] In the online Appendix, we establish the equivalence of $\hat{\beta}^{DID}_{s,s'}$ and $\hat{\beta}$ in the case of two states, and derive the expected value of $\hat{\beta}^{DID}_{s,s'}$.

[6] Because the effect of insurance β_s does not vary with time, the heterogeneous model satisfies the stable treatment effect assumption of de Chaisemartin and D'Haultfœuille (2018). Because the state s' is unaffected by Medicare, state s' satisfies the stable group assumption of de Chaisemartin and D'Haultfœuille (2018). Theorem 1 of de Chaisemartin and D'Haultfœuille (2018) implies that, under other standard conditions, the average of $\hat{\beta}^{DID}_{s,s'}$ is centered on the average coefficient among states affected by the policy change.

a two-way fixed effects estimator, thus allowing the researcher to estimate averages of the coefficients β_s.

Suggestions for Further Reading

Recently there has been a surge in interest in the role of treatment effect heterogeneity in the sorts of settings we discuss here, where policies are introduced with different intensities, or at different times, to different units. This is a very active area and it is not our intention to survey it fully. However, we can point to some published or forthcoming articles that readers may find helpful.

De Chaisemartin and D'Haultfœuille (2018) consider a setting closely related to the one we discuss here. They consider the possibility that treatment effects vary by unit and over time, and formalize issues that can arise with exposure-adjusted difference-in-differences estimators. They propose two alternative estimators, one of which corrects the exposure-adjusted difference-in-differences estimator directly for diverging trends due to differential exposure. De Chaisemartin and D'Haultfœuille (2020) extend the analysis to a more general setting with multiple time periods, and again propose a time-corrected difference-in-differences estimator that can help avoid issues of the sort we illustrate above.[7] The Stata packages *fuzzydid* and *did_multipledgt* implement both alternative estimators. Related to de Chaisemartin and D'Haultfœuille (2020), Imai and Kim (2020) characterize the relationship between a two-way fixed effects estimator and the difference-in-differences estimator, and use this to illustrate some pitfalls of the two-way fixed effects estimator.

A related but distinct setting is one of staggered adoption, where different units (for example, US states) adopt a policy (for example, unilateral divorce) at different times. In this setting, when policy effects may differ over time or across units based on when they adopt the policy, the two-way fixed effects estimator experiences issues similar to those we illustrate above. Goodman-Bacon (2021) proposes diagnostics for the performance of a two-way fixed effects estimator in such situations. The Stata package *bacondecomp* implements these diagnostics. Sun and Abraham (2021) propose an estimator that avoids some of the drawbacks of the two-way fixed effects estimator by taking advantage of the presence of never-treated units in the sample. The Stata package *eventstudyinteract* implements this estimator. Callaway and Sant'Anna (2021) propose a similar estimator that uses not-yet-treated units as control and can efficiently adjust for covariates using approaches developed in Sant'Anna and Zhao (2020). The Stata package *csdid* implements this estimator. Athey and Imbens (2022) consider the interpretation and variability of the difference-in-differences estimator in situations in which a unit's date of adoption is randomly assigned.

[7]Both articles by de Chaisemartin and D'Haultfœuille (2018, 2020) include applications to an earlier paper of Shapiro's (Gentzkow, Shapiro, and Sinkinson 2011). So, Shapiro is here to take advice as well as give it.

■ *We thank our dedicated research assistants for their contributions to this project. We are grateful to Clément de Chaisemartin, Amy Finkelstein, Andrew Goodman-Bacon, Ethan Lewis, Dan Levy, Pedro Sant'Anna, Matt Taddy, two anonymous reviewers, participants at the University of Michigan Two-Way Fixed Effects Econometrics Workshop, and the editors, Timothy Taylor and Heidi Williams, for helpful comments and suggestions. Liyang Sun gratefully acknowledges support from the Jerry A. Hausman Graduate Dissertation Fellowship and Institute of Education Sciences, U.S. Department of Education, through Grant R305D200010. Jesse Shapiro gratefully acknowledges support from the Eastman Professorship, the Population Studies and Training Center, and the JP Morgan Chase Research Assistantship, all at Brown University, and from the National Science Foundation under Grant No. 1949047.*

References

Athey, Susan, and Guido Imbens. 2022. "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption." *Journal of Econometrics* 226 (1): 62–79.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–30.

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

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

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2018. "Fuzzy Differences-in-Differences." *Review of Economic Studies* 85 (2): 999–1028.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.

Dube, Oeindrila, and Juan F. Vargas. 2013. "Commodity Price Shocks and Civil Conflict: Evidence from Colombia." *Review of Economic Studies* 80 (4): 1384–421.

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

Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. 2011. "The Effect of Newspaper Entry and Exit on Electoral Politics." *American Economic Review* 101 (7): 2980–3018.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–77.

Heckman, James J., and Edward J. Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments." In *Handbook of Econometrics*, Vol. 6, edited by James J. Heckman and Edward E. Leamer, 4875–5143. North Holland: Elsevier.

Imai, Kosuke, and In Song Kim. 2020. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29 (3): 1–11.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.

Nunn, Nathan, and Nancy Qian. 2011. "The Potato's Contribution to Population and Urbanization: Evidence from a Historical Experiment." *Quarterly Journal of Economics* 126 (2): 593–650.

Pierce, Justin R., and Peter K. Schott. 2016. "The Surprisingly Swift Decline of US Manufacturing Employment." *American Economic Review* 106 (7): 1632–62.

Sant'Anna, Pedro H.C., and Jun Zhao. 2020. "Doubly Robust Difference-in-Differences Estimators." *Journal of Econometrics* 219 (1): 101–22.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–199.

Zhang, Xiaoquan (**Michael**)**, and Feng Zhu.** 2011 "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia." *American Economic Review* 101 (4): 1601–15.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 205–220

Retrospectives
Sadie T.M. Alexander: Black Women and a "Taste of Freedom in the Economic World"

Nina Banks

This feature addresses the history of economic terms and ideas. The hope is to deepen the workaday dialogue of economists, while perhaps also casting new light on ongoing questions. If you have suggestions for future topics or authors, please contact either Beatrice Cherrier, CNRS & CREST, ENSAE-Ecole Polytechnique (beatrice.cherrier@gmail.com) or Joseph Persky, University of Illinois at Chicago (jpersky@uic.edu).

Introduction

In 1935, the nation's first African American PhD economist, Sadie Tanner Mossell Alexander, testified before the Pennsylvania legislature in opposition to a bill that purported to provide additional labor market protections to gainfully employed women. By then, most states had legislation that regulated women's hours of paid work in industry (Smith 1937; Goldin 1988). The legislation did not mention race explicitly, but its effect was to benefit white women workers while excluding nearly all African American women workers. Alexander stated (as quoted in Alexander 2021, p. 164):

> Recently I appeared before the House Committee of our Legislature to oppose House Bill No. 371 that proposed to decrease from fifty-four hours to forty hours labor by women in Pennsylvania except domestic servants

■ *Nina Banks is Associate Professor of Economics, Bucknell University, Lewisburg, Pennsylvania and Visiting Fellow, Institute on Race, Power and Political Economy, The New School, New York City, New York. Her email address is nina.banks@bucknell.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.205.

> and those employed in agriculture. Ninety-two percent of all Negro women in Pennsylvania gainfully employed are domestic servants.

Discrimination against Black women was not a new subject for Sadie Alexander. Born Sadie Tanner Mossell into a prominent family in Philadelphia, Pennsylvania, in 1898, she came of age during the progressive era of social reform that spanned the 1890s to the 1920s. When she received her PhD in economics in 1921 from the University of Pennsylvania, no academic institution was willing to hire her as an economist. Alexander worked for two years in North Carolina as an assistant actuary for an all-Black insurance company and then returned to Philadelphia in 1923 to marry attorney Raymond Pace Alexander. In Philadelphia, she could not get a job as a high school teacher because the city restricted employment in public secondary schools to white women (Alexander 2021). After spending an intellectually unsatisfying year at home, Alexander went back to the University of Pennsylvania and received her law degree in 1927.

Interest in the loss to the economics profession from Alexander's inability to practice economics began two years after she died, when Malveaux (1991, p. 308) explored the "missed opportunity for her, for the economics profession, and for the body of economic knowledge that pertains to African Americans." In 2003, when I began researching Alexander's archival records, the general belief was that Alexander had not retained much interest in economics after becoming an attorney (Malveaux 1991). However, I found that Alexander had left behind a treasure trove of speeches on the economic status of African Americans from the 1920s through the 1970s. Clearly, Alexander had continued to practice economics in the public domain throughout her life (Banks 2005, 2008).

Alexander's speeches and writings covered a range of topics in macroeconomics, labor, and economic history in relation to African American men and women; therefore, we should be careful to avoid delimiting her work as related to Black women. As one example, in her doctoral dissertation, "The Standard of Living Among One Hundred Negro Migrant Families in Philadelphia," she focused on the sudden increase in the number of Black people migrating to Philadelphia in response to World War I (Mossell 1921). Ten percent of the 400,000 African American migrants who left the South between 1916 and 1918 migrated to Philadelphia. In structured interviews with 100 Black migrant families, Alexander sought to determine the extent to which southern migrants could adapt to urban life in Philadelphia based upon their ability to earn a "fair standard of living."

In this essay, I will focus on Alexander's views about the labor market status of African American women. Indeed, Alexander's speeches and writing during the 1930s represent the first body of scholarship on Black women and work by an economist. Discussions of the history of women and work by US economists primarily focus on the experience of white women who entered the mills in the late eighteenth century and then left factories for domesticity by the mid-nineteenth century before returning to the labor force in large numbers during the 1970s. But since the late nineteenth century, African American women—regardless of their marital

status—have had higher labor force participation rates than other groups of US women, while lagging behind white women in terms of earnings (Jones 1985).

I will highlight several themes of Alexander's thought concerning Black women and work, noting ways in which her thinking was in opposition to prevailing views of women's economic roles. At a time when progressive attitudes about women in the workplace leaned heavily toward a belief that women should fulfill a moral role in the household rather than an economic role in the workplace—which often led to a belief that Black and other working class women were immoral for working—Alexander supported the idea of women in the workforce as a positive good for themselves, their families, and the economy. At a time when nearly all African American women in the workforce were either in domestic service or in agriculture, Alexander embraced the idea of women working in industrial jobs. Further, she argued that Black women's greater economic independence would enable them to exert greater political influence. Finally, at a time of Jim Crow segregation and racial violence, Alexander shed light on the racial contradictions embedded in New Deal policies that prevented African American women from receiving protections as workers and as mothers.

Progressive Era: Diverging Circumstances of Black and White Women

I begin with an overview of prevailing views on women's gainful employment that shaped the social reform efforts of African American and white women during the progressive era. White women generally experienced improvements in their status during the progressive era, especially after the ratification of the Nineteenth Amendment in 1920 granted women the right to vote, giving them a greater say in shaping public policies (Gordon 1991). Women industrial workers challenged low wages and unsafe working conditions by organizing and waging strikes through labor unions that were segregated by gender and, often, race (Barrett 1999). However, African American women experienced worsening of their circumstances; indeed, Nielson (1977) described this era as a "nadir" in relations between Black and white Americans because of the prevalence of Jim Crow segregation and racial terror.

Moreover, during this time, men and women also experienced public pressure to conform to newly developed expectations about gender roles. During the nineteenth century, white women's economic status had gone through changes that left them dependent on wage-earning men. White women and girls had provided the labor needed in the early mills and factories in the Northeast, but by the mid–nineteenth century, white male workers became the primary labor force in the growing industrial sector (Kessler-Harris 1982). Contemporary economist Edith Abbott (1910, p. 323) bluntly characterized the social pressure on women to first enter and then exit the workforce as stemming from men's economic interests:

> In the days when the earliest factories were calling for operatives the public moralist denounced her for "eating the bread of idleness," if she refused

> to obey the call. Now that there is some fear lest profuse immigration may give us an oversupply of labor, and that there may not be work enough for the men, it is the public moralist again who finds that her proper place is at home and that the world of industry was created for men.

The growth of a middle-class, professional workforce did enable middle-class men to support their wives financially at home. Victorian ideals of "true womanhood" developed, which viewed "good," proper women as moral caretakers of members of their households while economically dependent on men who fulfilled the role of breadwinners. However, this nineteenth-century separation of men and women into distinct realms of the private household sphere and the public sphere beyond the household never reflected the experiences of African American and other racialized women; indeed, it was more ideology than reality in the experiences of many working-class white women as well (Pitts 2014; Glenn 1985).

Nonetheless, by the late nineteenth century, white women who were active in progressive era social reform efforts—Edith and Grace Abbott, Jane Addams, Sophonisba Breckinridge, and others—embraced aspects of the domestic ideal. They expanded their own influence into the public domain by arguing that women's roles as moral guardians of family life extended beyond the private household to the public sphere in matters governing family welfare: child labor, alcoholism, sanitation, poverty, and public education. Moreover, efforts by white social reformers to ban "homework" done by working class women and children in tenements often embraced the domestic ideal in order to protect working class women from exploitive working conditions, appealing to women's roles as mothers and thus reinforcing their dependent economic status (Boris 2014, p. 74). State-level regulation of women's work was premised on the grounds that physical labor had an adverse impact on women's bodies. In 1908, the Supreme Court in *Muller v. Oregon* (208 US 412) upheld the right to limit women's hours of work: since "healthy mothers are essential to vigorous offspring, the physical well-being of a woman becomes an object of public interest and care in order to preserve the strength and vigor of the race."

The concern that white reformers had for protecting women's ability to produce healthy babies and provide care for children within their own homes did not extend to African American women. African American mothers, in particular, had higher labor force participation rates than their white counterparts. African American women's overall labor force participation rate was around 40 percent from 1870 to 1900, compared to a rate below 15 percent for white women (Boustan and Collins 2014). In Goldin's (1977) sample of seven southern cities in 1880, 35.4 percent of married Black women were in the labor force, compared with only 7.3 percent of married white women. Thistle (2006, p. 27) argues that African American women's workloads were so great that it was difficult for them to sustain their families: one-third of African American children died before reaching age ten and their mothers often died before their youngest child left home.

Moreover, because of occupational exclusion based on race and gender, about 90 percent of Black women in the labor force in the early 1900s worked in domestic

service or agriculture—sectors with little union representation. White families who hired Black women as domestic workers often subjected them to arbitrary work rules that included long hours with low and irregular pay. Victoria S., a live-in domestic worker during the early twentieth century, complained that she did not have regular work hours, never knew what days she was "allowed off" and "wasn't getting much" money. However, she continued to work for the family since "at that time, if they wanted you to work every day, you worked every day or you didn't have a job" (as quoted in Gottlieb 1976). The pay rate of Black women domestics in the early twentieth century, at $8 per month, was so low that the majority of poorest white families could afford Black domestic labor, therefore lessening the domestic responsibilities of white wives (Amott and Matthaei 1991).

During this time, the majority of Black wives were not employed in a job for pay. They were also more likely to have been married by the age of 35 compared to white women (Elliott et al. 2012). Nonetheless, because Black women were more likely to work outside of their households than white women were, they were subjected to public ridicule as bad women and mothers (Palmer 1983). White Americans hired African American mothers as workers without regard to their caregiving needs at home, and then demonized them as neglectful mothers because of their paid work.

Black women social reformers of this time addressed similar issues as white reformers—temperance, public education, voting, and care for the sick—while also challenging conditions of racial exclusion, segregation, inadequate public funding, and lynchings and mob violence that were directly relevant for Black Americans. However, prominent African American women reformers of this time like Anna Julia Cooper and Mary Church Terrell also promoted the importance of Black women as moral guardians of their households. Cooper (1892, p. 133) believed that Black women's power was in their morality such that when they "follow[ed] the instincts of [their] nature . . . [they] must always stand for the conservation of those deeper moral forces which make for the happiness of homes and the righteousness of the country." Mary Church Terrell (1898), a wealthy founding member of the most important African American women's social reform organization, the National Association of Colored Women (NACW), put the matter of Black women's morality in historical context:

> To the women of the race may be attributed in large measure the refinement and purity of the colored home. The immorality of colored women is a theme upon which those who know little about them or those who maliciously misrepresent them love to descant. Foul aspersions upon the character of colored women are assiduously circulated by the press of certain sections and especially by the direct descendants of those who in years past were responsible for the moral degradation of their female slaves.

Black social reformers were more likely to recognize the economic necessity of work for Black women. However, they also placed an emphasis on the morality of Black women in order to counter their public image as immoral and sexually

available and to shield them against the sexual harassment and assault that they experienced from white employers while working as domestic servants and farm laborers.

Alexander and Black Women's Industrial Work in the Progressive Era

In Alexander's first field research project while a graduate student, the Consumers' League of Eastern Pennsylvania employed her as a special investigator from September 1919 to June 1920, along with another young African American woman, to assist their white researchers with interviewing a sample of 190 Black women and girls recently hired by Philadelphia industries about their backgrounds, training, and work experiences (Consumers' League 1920, p. 11).[1] The Eastern Philadelphia League was part of the National Consumers League, which was formed in 1899 by middle- and upper-income, white women social reformers in response to the growth of consumer goods produced within factories rather than within households (Wolfe 1975). Organizers such as general secretary Florence Kelley (1901) sought to increase women's influence as consumers over product quality and safety along with working conditions and earnings for women factory workers. According to the Consumer's League of Eastern Pennsylvania (1920, p. 7), no industries in Philadelphia had hired Black women to any meaningful degree prior to World War I. However, the labor shortage and increased demand for workers during World War I prompted Philadelphia industries to hire African American women and girls.

From 1910 to 1919, the Black population in Philadelphia increased by 48 percent, consisting of 125,000 people. Quoting from local newspapers, the Consumers' League (1920, p. 8) reported that "Negro women are repairing railway tracks, making explosives, and serving as porters and inspectors in many industries here, taking the places of men who have gone to war or have entered other industries." In some industries, Black women were hired to replace men by "doing work that no white woman would do," according to one employer, because of the harshness of working conditions (Consumers' League 1920, p. 10).

The Consumers' League (1920) was interested in the increase in Black women's industrial employment, their pay rates, and employer perceptions about the caliber of Black women's work relative to white workers. Although some Black women and girls had worked as domestic servants in Philadelphia prior to the war, others were migrants from southern states. These women and girls worked in 28 Philadelphia

[1] At the same time, Alexander was separately conducting field research for her dissertation, but, as noted earlier, that project focused on interviews with 100 Black migrant families rather than on Black women in industrial jobs. When she briefly discussed women in her dissertation, it was to examine the contribution of mother's income to family income by noting the effect that caring for children had on women's employment. For example, she found that although 52 percent of migrant mothers in her study were wage earning, migrant wives without children and those with older children were more likely to supplement family income through employment than mothers in small families with young children (Mossell 1921, p. 16).

industries in a wide variety of positions that included cleaners at railroads and glass factories, machine operators at garment and knitting factories, and packers at paper novelty and tobacco companies (Consumers' League 1920, p. 12).

In some industries, such as department stores, Black women had worked at least ten years prior to the start of the war primarily in occupations where they did "not come in contact with the public" (Consumers' League 1920, p. 18). Employers paid Black women lower wages than white workers who occupied the same jobs. In other industries, such as garments, most Black and white women worked separately either because they worked in different occupations or because white women objected to working alongside Black women. One employer noted that they would discontinue hiring Black women because doing so made it "impossible to secure the best class of white girls if they had colored" (Consumers' League 1920, p. 41). In deference to white women, some employers segregated their bathrooms by race or prevented Black women from having access to the sole restroom reserved for white women. The Consumers' League (1920, p. 40) described Black women's working conditions compared to those of white women:

> They are often given the oldest and hardest machines to operate; they are not permitted to do piece work, which would increase their earnings; they are not allowed to work after they have earned a certain amount; in one case they have to go downstairs for their materials, while the white girls have materials brought to them; they have the darker and more poorly ventilated section, the smallest locker space, the worst sanitary provisions; and they are discriminated against in the matter of rest rooms.

Although we do not have direct information on Sadie Alexander's role in the Consumers' League Report beyond her work in interviewing Black women, she would have heard dozens of first-hand accounts from Black women workers in Philadelphia about their experiences. In the 1930s, Alexander wrote and gave public addresses on the status of Black women in the industrial sector.

The Case for Expanded Industrial Work by Black Women

In 1930, Alexander published an article in the National Urban League's *Opportunity Magazine: Journal of Negro Life* on "Negro Women in Our Economic Life."[2] She had initially presented the article as a paper at an Urban League conference in Buffalo, New York. Unlike other analyses, such as Abbott's (1910) *Women in Industry*, that focused on white women's experiences in the labor market, Alexander

[2]The National Urban League had originally formed in 1910 as the Committee on Urban Conditions Among Negroes, in response to the needs of African Americans who had migrated to urban areas in the North. Alexander's dissertation research had focused on African American migrants from the South to the North, and she began serving on the board of the National Urban League in New York in 1929.

discussed African American women workers. This short, incisive article placed the work of African American women at the center of analysis as well as provided an explanation of the impact of the rise of market relations on women's home production of goods and services.

Alexander examined three major effects of industrial work on women: the impact of the transition to a market economy on women's home production, the implications of increased participation in paid labor for Black women, and the benefits of Black women's participation in industrial work for Black women and families as well as to the nation.

To develop the first theme, Alexander noted that prior to the Industrial Revolution, men and women were joint producers and consumers in the family economy, such that the contribution of each to society was equally valued (Alexander 2021). However, as production that had traditionally been performed at home came to be produced outside of the household, production in the new social order became associated with producing commodities or providing services with a money price. Because women's home responsibilities did not have a market price, their services were regarded as valueless (Alexander 2021, p. 53):

> Modern industrial processes, having robbed the home of every vestige of its former economic function, left in the home to be performed by the woman only those services which are as "valueless" and "priceless" as air and water but not recognized as *valuable* in a price economy, where standards of value are money standards.

Alexander argued that in order for women to become producers rather than primarily consumers in the Industrial Age, they needed to become industrial workers outside of their households.

With regard to the second theme, Alexander (2021, p. 55) advocated that women seek to change their status from homemaker to more skilled and better paid industrial workers and produce goods that had a recognized price value. She had no illusions that this would be an easy transition for Black women; for example, a theme that overlapped with the Consumers' League report and Alexander's later speeches in the 1930s was the relegation of Black women to many of the worst jobs within industry and their marginal status as "the last to be hired and the first to be fired"—a problem that hampered their ability to have higher earnings and mobility (Alexander 2021, p. 112). However, Alexander observed in her 1930 article that women were in fact making this transition to industrial work through their increased labor force participation, from 1.3 million in 1870 to 7.3 million in 1920, and further noted that Black women had already been working outside of their households to a greater extent than white women. In 1910, for instance, 54.7 percent of Black women age 10 and over were gainfully employed compared to 19.6 percent of white women in the same age group, and 38.9 percent of Black women age 10 and over were employed compared to 16.1 percent of white women in same age group in 1920.

Despite their higher rates of participation, Black women were at the bottom of the hierarchy in both pay and occupations compared to men and white women, in substantial part because they were principally employed in domestic service and farming.

Alexander's discussion built to the third theme: participation of Black women in industrial work created benefits for Black women and families as well as for the nation. She stated that working in industry enabled Black women to have shorter working days than their other labor-market alternatives, so they would have more leisure time during which they could pursue industrial training classes in the evenings. Working in the industrial sector, she believed, also afforded Black women a degree of self-respect and social intercourse that was lacking in domestic service. Alexander countered the prevalent belief that family life suffered with wives' employment by saying that when women were valued in their work, they became happier, and that this would invariably improve home and family life (Alexander 2021, p. 55).

Alexander also maintained that Black women's increased labor supply would help to reduce the cost of producing output, therefore benefitting consumers through lower prices and increased consumer choices of goods. She argued that mass production would not have occurred in some industries without the additional labor supply of Black women (Alexander 2021, p. 57):

> Within the two decades during which Negro women have entered industry in large numbers, production has increased at such a rapid rate that economists have been forced to change their theory of a deficit economy, based on the assumption that population would always press upon food supply, to a theory of surplus economy. While the labor of Negro women cannot be held as the efficient cause of the mass production, it is submitted that without this available labor supply at a low price, mass production in many industries would not have been undertaken.

Alexander's discussion of the reclassification of women's activities, as the work that they previously did at home came to be performed outside of the household for a price, corresponds to a central argument made by feminists in the late twentieth century. However, later feminist economists often sought to elevate the unpaid work performed within the household by arguing that it had great social value. In contrast, Alexander inverted the prevailing dominant gender perspective that placed value on homemakers and disparaged wage-earning women by insisting that it was wage-earning women—not homemakers—who created value in the market-based economy. Viewed from the standpoint of race, Alexander was making a daring claim that placed value on the work of employed Black women—including mothers—over that of white homemakers.

African Americans of that time often expressed ambivalence about Black mothers' employment because it could be viewed as demonstrating the inability of Black men to fulfill the socially desired role of breadwinner (Harley 1990). To white

Americans, Black women's employment was a sign of racial backwardness (Bederman 1992). Alexander, however, provided a full endorsement of Black women's gainful employment. Her views on wives' employment were consistent with that of other Black American women reformers in the early twentieth century. Whereas most white reformers until World War II viewed married women's employment as a "misfortune" that was detrimental to women as well as to men and children, Black reformers tended to be more accepting of married Black women's employment and economic independence because they knew that, unlike white women, Black women were likely to be in the labor force for much of their lives (Gordon 1991, pp. 468, 470).

Views on mothers' employment also help to explain the different emphasis that Black and white reformers placed on childcare centers. Most white women reformers did not advocate for childcare centers—called "day nurseries" prior to the 1930s—while Black women reformers had engaged in fundraising activities to support day nurseries and kindergartens to meet the needs of gainfully employed Black women in the late nineteenth century (Jones 1982; Gordon 1991). In 1902, the National Federation of Day Nurseries recognized only 250 day nurseries in the country, resulting in wage-earning women either leaving children at home alone, taking them to their workplaces, or relying on families, friends, and neighbors for childcare (Durst 2005, p. 141). Alexander (1938) was a proponent of nursery schools in part because they met mothers' needs, but she also believed that day nurseries would serve children's needs better than parents who lacked training in infant care.

Moreover, Alexander did not share the concern of white reformers that women's economic independence would lead to women's sexual independence and the undermining of family and community morals (Feldstein 2018). White reformers advocated for a family wage that enabled white men to be breadwinners and their wives to be homemakers (Boris 2014). Alexander, however, believed that women needed to have their own income so that they were not dependent on men. Alexander delivered another speech during the 1930s, titled *Emancipated Woman*, where she elaborated on the implications of women's employment on family life and on the position of women within the political economy (Alexander 2021, p. 59):

> Women engaged in labor, postponed marriage, or controlled births. Those who have families permitted them to rear themselves. Divorces increased—not because the contact with [the] world made women less faithful, but because she no longer had to put up with embarrassments, neglects, and cruelties in order to secure bread and butter—she could make that for herself. Having gained a taste of freedom in the economic world, naturally women began to demand equality in the political world in order, if for no other reason that they might help legislate for their own protection.

Alexander reasoned that with "a taste of freedom in the economic world," women would become less economically dependent on men and more likely to

seek additional political freedoms. She believed that women should use their votes wisely by showing "tangible evidence of making this a better world to live in" (Alexander 2021, p. 61). Thus, Alexander's strategy for the full emancipation of African American women involved both earning an income to "be independent in her thought and action" and using political freedom to secure rights and equal treatment for African Americans (Alexander 2021, p. 64). Alexander also advocated in her speeches for white workers to recognize their commonalities with Black workers through collective bargaining in order to diminish racial antagonisms and improve overall earnings.

The Exclusion of Black Women from New Deal Protective Legislation

Alexander's admonition that women should use their votes to improve the lives of other Black Americans was consistent with increased political organizing by both Black and white women after gaining the vote in 1920. By the 1930s, with the Great Depression underway, white women social reformers including Mary W. Dewson, Frances Perkins, and Edith and Grace Abbott endeavored to shape New Deal legislation so that it would reflect their social welfare priorities (Jabour 2021).

However, compared to white women who had personal and professional connections to powerful white men, Black women had fewer opportunities to exert influence on New Deal policies (Gordon 1991). Sadie Alexander had hoped that the Roosevelt administration would enact policies upon taking office in 1933 that benefitted Black workers by ensuring that they would have equal access to labor market protections. Instead, New Deal policies typically excluded the occupations where the majority of African American women worked—agriculture and domestic service—from receiving benefits and labor protections. Alexander became a fierce critic of New Deal policies, because she believed that they had worsened the economic position of Black workers relative to white workers through exclusions and unfair administration of federal aid.

In a speech called "The Economic Status of Negro Women: An Index to the Negro's Economic Status," Alexander assessed the overall position of the Black community based on the position of Black women across sectors of the workforce (Alexander 2021). Black women's economic status was so vital to the wellbeing of family life that Alexander regarded it as a barometer of the status of the Black community in general (Alexander 2021, p. 65):

> I have been asked to talk to you about the Economic Status of the Negro Woman. While I shall chiefly confine my remarks to that subject, I am quite certain you will immediately realize that a picture of the economic status of our women is an index of our entire economic life. The hackneyed but nevertheless trite saying that a race can rise no higher than its women, applies to the economic status of a race as well as its moral and intellectual standing. So that a nation of women that forms, as do our women, the

> marginal workers in a pitifully small number of industries and the bulk of the domestic servants necessarily indicates the reduced economic status of the entire race.

Again, Alexander emphasized women's importance not simply as moral contributors, but also as economic contributors.

Alexander believed that high rates of employment for Black women during the Great Depression should be viewed as an indication of the toll that the Depression was taking on Black families because Black women were working out of necessity, given the high rates of unemployment for Black men. Additionally, of the nearly 40 percent of Black women over the age of 10 who were employed in 1930—twice the rate of white women—over 60 percent were in domestic service, 27 percent in agriculture, and less than 6 percent in industrial jobs, where they primarily worked in unsanitary jobs rather than as machine operators (Alexander 2021, p. 67). Black women's employment in domestic service and agriculture meant that they were especially vulnerable to job loss during the Great Depression, so Alexander's speech discussed the effects of Black women's job loss on families. Loss of jobs for both Black women and men, she stated, put a "great strain" on Black families (Alexander 2021, p. 69).

Alexander's views on the necessity of employment of Black wives during the Depression was at odds with public sentiment that maternal employment led to inadequate maternal supervision and moral guidance, which in turn led to juvenile delinquency (Pickett 2017). For example, contemporary sociologist Louise V. Kennedy (1930) stated, "Due to their low economic status, Negroes show an abnormally high percentage of married women engaged in gainful occupations, and the outside employment of mothers has a direct effect upon the amount of juvenile delinquency." Public sentiment also discouraged maternal employment as contributing to job loss for men and eroding their authority as breadwinning heads of households. In a 1936 Gallup poll, 72 percent of respondents stated that they disapproved of married women's employment if their husbands could support them (Saad 2017). Frazier's (1939) *The Negro Family in the United States*, the most prominent analysis of African American families published during the 1930s, stressed the importance of "traditional" gender roles with male providers and economically dependent wives as a foundation for stable family life. However, most Black men during the Depression were not in a position to support their wives. According to Feldstein (2018), most studies of Black and white families during the Great Depression defined "normal" families as white families that continued to rest on male authority such that they would be able to withstand unemployment and temporary receipt of public assistance.

Alexander (2021, pp. 69-70) also denounced Black women's exclusion from New Deal minimum wage protections.

> Think of the ridiculous position occupied by the Negro servant. Her employer, we will say, is a bank clerk. His salary and hours of work are protected by a code which was scientifically established by members of the

> Brain Trust. They determined in arriving at the proper salary to be paid the bank clerk the average cost of a proper amount of insurance protection. They made allowances for recreation and leisure, so as to induce spending. These learned gentlemen decided that a minimum of $45.00 per week is necessary for a fair standing of living for our Negro servant's bank clerk employer. Then they suddenly realize that his wife must have a maid or else she cannot leave the children at home in the evening, when the money provided for leisure and recreation is to be spent. So they add the large sum of $5.00 for a maid, and fix the minimum salary for the bank clerk at $50.00. I ask you; does not the maid have to carry insurance? Does she too not require recreation and leisure? Is she not also entitled to a place she may call home and some food when she goes there? Five dollars per week is ample to supply her needs—say the experts.

Notice that in this excerpt, Alexander compared the earnings of Black women to that of white men because of Black women's importance as financial contributors to family income. It was a recognition that African American women were often co-breadwinners within their households.

In their analysis of New Deal protective legislation for women, Mutari, Power, and Figart (2002, p. 55) argue that the exclusion of Black women from minimum wage and maximum hours legislation indicated that the state did not recognize Black women's needs as mothers as they did for white women, nor as paid laborers as they did for white men. In this sense, the New Deal protective legislation reinforced Black women's subordinate economic status as well as the prevailing racialized notions of homemaker and breadwinner.

The exclusion of African American women from the New Deal's major social welfare program—the 1935 Social Security Act—illustrates the dilemma that Black women faced as socially devalued workers and mothers. As domestic and farm workers, they were excluded from the unemployment and old age provisions that were designed to benefit male breadwinners. They also faced exclusions from the Aid to Dependent Children program, because cash assistance was administered at the state level in a manner that was discriminatory towards Black mothers. White social reformers who drafted the program had argued in favor of tying receipt of benefits to the mother's moral character and, indeed, state provisions included the mother's moral character as a factor in determining eligibility (Feldstein 2018). White caseworkers often denied benefits to Black mothers on the basis that they should be gainfully employed as domestic servants or because they deemed them to be morally undeserving.

Conclusion

White social reformers of the progressive era provided a vision of women's emancipation that often emphasized women's morality and reliance on male

breadwinners. In contrast, Sadie Alexander's views on women's roles within the economy rested on the labor market experiences of African American women. She argued for an expansion of women's rights to gainful employment and economic independence based on a belief that women must have their own earnings to gain economic independence and political freedom. Examining the differences between Alexander's views and that of her contemporaries should encourage modern economists to think critically about the long legacy of racialized notions of breadwinner and homemaker that have continued to shape public priorities governing labor protections as well as the rules regarding provision of childcare and cash assistance to poor mothers and children.

References

Abbott, Edith. 1910. *Women in Industry; A Study in American Economic History*. New York: D. Appleton and Company.

Alexander, Sadie Tanner Mossell. 1938. Letter from Mossell to Miss Vera Burks, YWCA, Atlantic City, NY. University of Pennsylvania Archives, STMA, Box 3, FF 33.

Alexander, Sadie Tanner Mossell. 2021. *Democracy, Race, and Justice: The Speeches and Writings of Sadie T.M. Alexander,* edited by Nina Banks. New Haven and London: Yale University Press.

Amott, Teresa, and Julie Matthaei. 1991. *Race, Gender, and Work: A Multicultural Economic History of Women in the United States*. Boston: South End Press.

Banks, Nina. 2005. "Black Women and Racial Advancement: The Economics of Sadie Tanner Mossell Alexander." *Review of Black Political Economy* 33 (1): 9–24.

Banks, Nina. 2008. "The Black Worker, Economic Justice, and the Speeches of Sadie T.M. Alexander." *Review of Social Economy* 66 (2): 139–61.

Barrett, Nancy J. 1999. "The Struggles of Women Industrial Workers to Improve Work Conditions in the Progressive Era." *OAH Magazine of History* 13 (3): 43–49.

Bederman, Gail. 1992. "'Civilization,' the Decline of Middle-Class Manliness, and Ida B. Wells's Anti-lynching Campaign (1892–94)." *Radical History Review* 52 (winter): 5–30.

Boris, Eileen. 2014. "Reconstructing the 'Family': Women, Progressive Reform, and the Problem of Social Control." In *Gender, Class, Race, and Reform in the Progressive Era*, edited by Nancy S. Dye and Noralee Frankel, 73–86. Lexington: The University Press of Kentucky.

Boustan, Leah Platt, and William J. Collins. 2014. "The Origin and Persistence of Black-White Differences in Women's Labor Force Participation." In *Human Capital in History: The American Record*, 205–40. Chicago: National Bureau of Economic Research.

Consumers' League of Eastern Pennsylvania. 1920. *Colored Women as Industrial Workers in Philadelphia*. Philadelphia: Consumers' League of Eastern Pennsylvania.

Cooper, Anna Julia. 1892. *A Voice from the South: By A Black Woman of the South*. Xenia, Ohio: Aldine Printing House.

Durst, Anne. 2005. "'Of Women, by Women, and for Women': The Day Nursery Movement in the Progressive-Era United States." *Journal of Social History* 39 (1): 141–59.

Elliott, Diana B., Kristy Krivickas, Matthew W. Brault, and Rose M. Kreider. 2012. "Historical Marriage Trends from 1890-2010: A Focus on Race Differences." Social, Economic, and Housing Statistics Division Working Paper SEHSD-WP2012-12.

Feldstein, Ruth. 2018. "'The Women Have a Big Part to Play': Citizenship, Motherhood, and Race in New Deal Liberalism." In *Motherhood in Black and White: Race and Sex in American Liberalism, 1930–1965*,

12–39. Ithaca: Cornell University Press.

Frazier, E. Franklin. 1939. *The Negro Family in the United States*. Chicago: University of Chicago Press.

Glenn, Evelyn Nakano. 1985. "Racial Ethnic Women's Labor: The Intersection of Race, Gender and Class Oppression." *Review of Radical Political Economics* 17 (3): 86–108.

Goldin, Claudia. 1977. "Female Labor Force Participation: The Origin of Black and White Differences, 1870 and 1880." The Journal of Economic History, 37 (1): 87–108.

Goldin, Claudia. 1988. "Maximum Hours Legislation and Female Employment: A Reassessment." *Journal of Political Economy* 96 (1): 189–205.

Gordon, Linda. 1991. "Black and White Visions of Welfare: Women's Welfare Activism, 1890–1945." *Journal of American History* 78 (2): 559–90.

Gordon, Linda. 1999. *Pitied But Not Entitled: Single Mothers and the History of Welfare, 1890–1935*. Cambridge, MA: Harvard University Press.

Gottlieb, Peter. Interview with Victoria S. May 25, 1976. Archives of the Industrial Society. Southern Blacks Migration to Pittsburgh 1916–1977, Oral History Collection. University of Pittsburgh, Pittsburgh, PA.

Harley, Sharon. 1990. "For the Good of the Family and Race: Gender, Work, and Domestic Roles in the Black Community, 1880–1930." *Signs: Journal of Women in Culture and Society* 15 (2): 336–49.

Jabour, Anya. 2021. "'It's Up to the Women': Eleanor Roosevelt, Women's Activism, and Human Rights." National Park Service. https://www.nps.gov/elro/learn/historyculture/it-s-up-to-the-women.htm (accessed March 2022).

Jones, Barbara A.P. 1985. "Black Women and Labor Force Participation: An Analysis of Sluggish Growth Rates." *Review of Black Political Economy* 14 (2–3): 11–31.

Jones, Beverly W. 1982. "Mary Church Terrell and the National Association of Colored Women, 1896 to 1901." *The Journal of Negro History* 67 (1): 20–33.

Kelley, Florence. 1901. "The Consumers' League." *American Journal of Nursing* 1 (9): 646–49.

Kennedy, Louise Venable. 1930. *The Negro Peasant Turns Cityward: Effects of Recent Migrations to Northern Centers*. New York: Columbia University Press.

Kessler-Harris, Alice. 1982. *Out to Work: A History of Wage-Earning Women in the United States*. New York: Oxford University Press.

Malveaux, Julianne. May 1991. "Missed Opportunity: Sadie Tanner Mossell Alexander and the Economics Profession." *American Economic Review* 81 (2): 307–10.

Mossell, Sadie Tanner. 1921. "The Standard of Living among One Hundred Negro Migrant Families in Philadelphia." PhD diss. University of Pennsylvania.

Muller v. Oregon, 208 U.S. 412 (1908).

Mutari, Ellen, Marilyn Power, and Deborah M. Figart. 2002. "Neither Mothers Nor Breadwinners: African-American Women's Exclusion From US Minimum Wage Policies, 1912–1938." *Feminist Economics* 8 (2): 37–61.

Nielson, David Gordon. 1977. *Black Ethos: Northern Urban Negro Life and Thought, 1890–1930*. Westport: Greenwood Press.

Palmer, Phyllis Marynick. 1983. "White Women/Black Women: The Dualism of Female Identity and Experience in the United States." *Feminist Studies* 9 (1),:151–170.

Pickett, Justin T. 2017. "Blame Their Mothers: Public Opinion About Maternal Employment as a Cause of Juvenile Delinquency." *Feminist Criminology* 12 (4): 361–83.

Pitts, Martha. 2014. "Nineteenth-Century Motherwork: Ideology, Experience, and Agency in Autobiographical Narratives by Black Women." In *Patricia Hill Collins: Reconceiving Motherhood*, edited by Kaila Adia Story. Bradford: Demeter Press.

Saad, Lydia. 2017. "Gallup Vault: A Sea Change in Support for Working Women." *Gallup*, July 20. https://news.gallup.com/vault/214328/gallup-vault-sea-change-support-working-women.aspx.

Smith, Florence P. 1937. *State Labor Laws for Women: Revision of Bulletin 98—Hours, Home Work Prohibited or Regulated Occupations, Seats, Minimum Wage*. Bulletin of the Women's Bureau, no. 144. Washington DC: United States Government Printing Office.

Terrell, Mary Church. 1898. "The Progress of Colored Women." Speech, National American Women's Suffrage Association fiftieth anniversary meeting, Washington DC, February 18, 1898.

Thistle, Susan. 2006. *From Marriage to the Market : The Transformation of Women's Lives and Work*. Berkeley: University of California Press.

Wolfe, Allis Rosenberg. 1975. "Women, Consumerism, and the National Consumers' League in the Progressive Era, 1900–1923." *Labor History* 16 (3): 378–92.

Journal of Economic Perspectives—Volume 36, Number 4—Fall 2022—Pages 221–228

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

Jonathan Portes has edited a collection of nine essays about *The Economics of Brexit: What Have We Learned?* (Center for Economic Policy Research, June 2022, https://cepr.org/publications/books-and-reports/economics-brexit-what-have-we-learned). From the overview by Portes: "However, as [Theimo] Fetzer points out, aggregate impacts are not the whole story by any means. His analysis suggests not only that the costs of Brexit are very unevenly distributed, but that, perhaps paradoxically, those areas that voted most heavily for Brexit are the worst affected, while London has escaped largely unscathed, at least so far." Portes also describes his own research: "I describe the new system, which does indeed represent a very significant tightening of controls on EU migration compared to free movement. . . . However,

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.36.4.221.

compared to the current system—and in contrast to earlier predictions—the new proposals represent a considerable liberalisation for non-EU migrants, with lower salary and skill thresholds and no overall cap on numbers. This implies that about half of all full-time jobs in the UK labour market could in principle qualify an applicant for a visa. This represents a very substantial increase—perhaps a doubling compared to the previous system—and also makes the new system considerably more liberal with respect to non-European migrants than that of most EU member states, which typically apply much more restrictive (de facto and/or de jure) skill or salary thresholds, and often enforce a resident labour market test. . . . So, the new system does not represent an unequivocal tightening of immigration controls . . ."

David Autor discusses "The labor market impacts of technological change: From unbridled enthusiasm to qualified optimism to vast uncertainty" (one of four essays in *An Inclusive Future? Technology, New Dynamics, and Policy Challenges*, edited by Zia Qureshi, Brookings Institution, May 2022, https://www.brookings.edu/research/an-inclusive-future-technology-new-dynamics-and-policy-challenges/). "[W]hat is the role of technology—digital or otherwise—in determining wages and shaping wage inequality? The answer is not obvious, and the successive evolution of thinking on this topic reflects the subtlety of the question. I present four answers below, corresponding to four strands of thinking on this topic, and discuss the distinct implications of each. I refer to these four paradigms as the education race, the task polarization model, the automation-reinstatement race, and the era of Artificial Intelligence uncertainty. The nuance of economic understanding has improved across each of these epochs. Yet, traditional economic optimism about the beneficent effects of technology for productivity and welfare has eroded as understanding has advanced. Given this intellectual trajectory, it would be natural to forecast an even darker horizon ahead. I refrain from doing that, however, because forecasting the 'consequences' of technological change treats the future as a fate to be divined rather than an expedition to be undertaken." He also estimates: "[M]ore than 60 percent of employment in 2018 was found in job titles that did not exist in 1940 . . . The introduction of new work, however, is not uniform across skill groups. Between 1940 and 1980, most new work that employed non-college workers was found in construction, transportation, production, clerical, and sales jobs—which are squarely middle-skill occupations. In the subsequent four decades (1980–2018), however, the locus of new work creation for non-college workers shifted away from these middle-tier occupations and towards traditionally lower-paid personal services. Conversely, new work creation employing college-educated workers became increasingly concentrated in professional, technical, and managerial occupations."

Richard Baldwin discusses "Globotics and macroeconomics: Globalisation and automation of the service sector" (presented at the ECB Forum on Central Banking 2022, June 27–29, https://www.ecb.europa.eu/pub/conferences/html/20220627_ecb_forum_on_central_banking.en.html, where videos of presentations and comments are included). "In a nutshell, digital technology (digitech) is rapidly exposing services that were previously non-tradeable to the opportunities and

challenges of globalisation. . . . Simultaneously, digitech is introducing automation to services that were previously non-automatable. 'White-collar robots' is one name for the automating algorithms—things like Robotic Process Automation (RPA), virtual assistants, chatbots, and sophisticated AI packages like IBM's Watson. . . . To stress that both the globalisation and robotisation of service jobs are happening at the same time—and are driven by the same technologies—I created the ugly, but hopefully memorable word 'globotics' in my 2019 book on the subject. In my view, globotics will improve lives in the long run but the transition could be rough. . . . Services are hard to tax at the border, so most barriers arise from domestic regulation (OECD 2020). Much of this regulation, however, concerns 'final' services, not 'intermediate' services. Regulations, restrictions, and controls typically apply only to transactions between the final service seller and the final service buyer. The service tasks that are inputs to these final services are—by contrast—much less regulated. For example, while there are strict rules for selling accounting services in the US, there are few rules concerning the qualifications of the service workers that do the paperwork behind the provision of such accounting services. A US accountant can employ pretty much anybody to tally up a client's travel expenses and collate them with expense receipts. The quality control burden falls on the sellers of the final service, not government regulators. . . . [D]igitech is rapidly lowering the technological barriers to trade in intermediate services. These two facts mean that service-trade barriers are falling radically faster than goods-trade barriers and likely to continue doing so for the foreseeable future. . . . [E]xport capacity in emerging markets is not as great a limiting factor in services as it is in goods since every nation has a workforce that is already producing intermediate-service tasks. All emerging market economies have bookkeepers, forensic accountants, CV screeners, administrative assistants, online client help staff, graphic designers, copyeditors, personal assistants, travel agents, software engineers, lawyers who can check contracts, financial analysts who can write reports, etc. There is no need to develop whole new sectors, build factories, or develop farms or mines. This fact is the basis of a broad re-evaluation of development pathways for emerging markets . . ."

Darrell Duffie and Elizabeth Economy have edited *Digital Currencies: The US, China, and the World at a Crossroads*, based on the discussions of a task force convened at the Hoover Institution at Stanford University (March 2022, https://www.hoover.org/research/digital-currencies-us-china-and-world-crossroads). "Central bank digital currencies (CBDCs) have taken flight globally. More than ninety central banks are researching, piloting, or deploying CBDCs. Several are already testing cross-border transactions. Among the countries exploring CBDCs, China occupies a particularly important position. It is the first major country to deploy a CBDC widely within its own economy, and its central bank is dominant among those participating in a cross-border payments development project under the auspices of the Bank for International Settlements. China's emergence as a first mover in this space gives Beijing a significant opportunity to cement its international leadership of payments technology innovation and adoption, to set economic norms and technical standards that align with its authoritarian governance system, and to increase its ability

to undercut the traditional dominance of the US dollar as a source of geoeconomic and strategic influence."

Distressed Labor Markets

Timothy J. Bartik offers some thoughts in "How State Governments Can Target Job Opportunities to Distressed Places" (Upjohn Institute Technical Report No. 22-044, June 2022, https://research.upjohn.org/up_technicalreports/44/). There's also an overview in *Employment Research Newsletter* from the Upjohn Institute for Employment Research (August 2022, https://research.upjohn.org/empl_research/vol29/iss3/1). From the newsletter: "Distressed places, which have low employment to-population ratios (employment rates), are a big problem in America. . . . About two-fifths of all Americans live in local labor markets whose employment rate for prime-age workers (ages 25–54) is more than 5 percentage points below full employment. For neighborhoods, about one-fifth of all Americans live in census tracts whose prime-age employment rate is more than 5 percentage points below their local labor market's average. These low employment rates are linked to major social problems: substance abuse, crime, and family stress. . . . Local job creation is most cost-effectively accomplished by providing businesses with 'customized services' such as infrastructure, customized job training, and business advice programs—including manufacturing extension services. Such customized services have less than one-third the cost-per-job-created of business tax incentives. In contrast, in a distressed neighborhood, more neighborhood jobs will not much help the neighborhood's residents, as most neighborhood jobs are not held by residents. Residents of distressed neighborhoods can best be helped by services to increase job access, including better transportation, job training, and child care. . . . Total annual costs for all states would come to $30 billion annually—$21 billion for local labor markets and $9 billion for neighborhoods. This $30 billion cost is affordable, as it is less than 3 percent of overall state taxes. Many states could cover the required costs by replacing their business tax incentives."

Cityscape, from the US Department of Housing and Urban Development, published an 11-paper symposium about "An Evaluation of the Impact and Potential of Opportunity Zones" (2022, 24: 1, https://www.huduser.gov/portal/periodicals/cityscape.html). In his introduction, Daniel Marcin writes: "Opportunity Zones allow investors with capital gains to reinvest that money into Qualified Opportunity Funds (QOF), which then invest in OZs. Doing so has three main benefits. 1. The capital gains tax due on the original investment sale is deferred until the sale of the QOF investment or the end of 2026, whichever comes first. 2. If the investor holds the QOF investment for 5 years, the cost basis of the investment is increased by 10 percent. If held for 7 years, or 2 additional years, the cost basis increases by an additional 5 percent. 3. If the QOF investment is held for 10 years, then no tax is due on any gains on the OZ investment (IRS, 2021a)." In their essay, Blake Christian and Hank Berkowitz argue: "The federal OZ program is arguably one of the most

flexible, impactful, and bipartisan tax programs for helping disadvantaged communities in half a century." They cite estimates that $75 billion had been invested in the opportunity zone program by the end of 2020.

Interviews with Economists

Noah Smith presents an "Interview: Leah Boustan, economist," with the subtitle, "In which we talk all about immigration" (Noahopinion, July 17, 2022, https://noahpinion.substack.com/p/interview-leah-boustan-economist). "Americans vastly overestimate how many immigrants are in the country today. According to a survey conducted by Stefanie Stantcheva and her co-authors, Americans guess that 36 percent of the country is born abroad, whereas the real number is 14 percent. So, this misconception gives rise to fears that we are in an 'immigration crisis' or that we have a 'flood' of immigrants coming to our shores. In reality, the immigrant share of the population today (14 percent) only just reached the same level as it was during the Ellis Island period for over 50 years! After this, I would say that the second biggest misconception is that immigrants nowadays are faring more poorly in the economy and are less likely to become American than immigrants 100 years ago. . . . We find that Mexican immigrants and their children achieve a substantial amount of integration, both economically and culturally. . . . [T]he pattern . . . whereby the kids of poor and working-class immigrants do better than their American counterparts, is true both today and in the past. The children of poor Irish or Italian immigrant parents outperformed the children of poor US-born parents in the early 20th century; the same is true of the children of immigrants today. We are able to delve into the reasons for this immigrant advantage in the past in great detail, and we find that the single most important factor is geography. Immigrants tended to settle in dynamic cities that provided opportunities both for themselves and for their kids. . . . Geography still matters a lot today, but not as much as in the past. Instead, we suspect that educational differences between groups matter today. Think about a Chinese or Indian immigrant who doesn't earn very much, say working in a restaurant or a hotel or in childcare. In some cases, the immigrant him or herself arrived in the US with an education—even a college degree—but has a hard time finding work in their chosen profession. Despite the fact that these immigrant families do not have many financial resources, they can pass along educational advantages to their children."

David A. Price interviews Stephanie Schmitt-Grohé in the most recent issue of *Econ Focus* (Federal Reserve Bank of Richmond, Third Quarter 2022, pp. 24–28, https://www.richmondfed.org/publications/research/econ_focus/2022/q3_interview). In describing recent research with Martín Uribe, she says: "We wanted to answer the question that I think everybody is interested in: Is this inflation hike temporary or permanent? Our idea was that during the postwar period—since 1955, say—the only big inflation was the inflation of the 1970s. . . . So we said, since the current inflation is unprecedented in the postwar period, what will we see if we just go

further back in history? Because we wanted to go back in history, we used the database of Òscar Jordà, Moritz Schularick, and Alan Taylor, which goes back to 1870. . . . We found that if we estimate the model since 1955, which is what most people do when they talk about cyclical fluctuations—actually, many people only start in the 1990s or look at the last 30 or 40 years, the so-called Great Moderation period—the model is led to interpret the entire current increase in inflation as permanent. But if the model is given the chance to look back further in time, where we had more episodes of a short-lived and large inflation spike, the interpretation is that only 1 or 2 percent of the current increase in inflation is of a more permanent nature. An example to look at is the Spanish Influenza of 1918 in the United States. That was also a period of an inflation spike, but inflation had started already a year or two before the influenza pandemic. There were similarities to now, namely a pandemic and high inflation. There was a small increase in the permanent component of inflation during the years around the influenza pandemic, but the majority of it was transitory."

Shruti Rajagopalan conducts a wide-ranging two-part interview with Lant Pritchett (*Discourse*, "Ideas of India," first part published on March 17, 2022, https://www.discoursemagazine.com/economics/2022/03/17/ideas-of-india-where-did-development-economics-go-wrong/; second part on June 9, 2022, https://www.discoursemagazine.com/culture-and-society/2022/06/09/ideas-of-india-reforming-development-economics/). "[M]ost of the way the regime for mobility of persons around the world has worked since the 1920s is that people who are allowed to work in a country are either citizens or on a path of citizenship in the country. I'm actually a big advocate of separating those two things and saying the needs of U.S. or Germany or France for labor are not being met. Because if the only way in which a person can come and work in France—to take care of the elderly or perform relatively low-skilled services—is by allowing that person to become a French citizen, the political consensus is no. We'll prefer not having the service. . . . My big thing is if we actually had rotational mobility, in which people could come and perform the labor services but not necessarily instantaneously be on the path to citizenship, this could be a big thing that would be a win-win-win. It would be a win for the countries that need the labor. It would be a win for the workers that move. It would be a win for the sending countries. . . . Doesn't sound like the world's getting friendlier to open borders. That said, the needs for this labor are going to get so huge, in my view, that there needs to be some intermediate solution. I think a well-regulated industry that does rotational mobility is a massive, massive opportunity."

Pritchett is also asked: "What is the role of an economist?" He answered: "I don't really teach undergraduates very often, but I was invited to give the opening lecture to a development economics course of undergraduates. My take was that economics is the social science of love. It's the truly loving social science, and what I meant—and they were, of course, like, 'What? Economics and love? That's crazy.'But think about what economists do. We take individuals—objective functions are objective functions. We don't start with any premise about what would be good for society or good for X or good for Y. But I think economists, when they're doing it right,

they start from, what is it that people want to accomplish with their lives? Okay. Let's think about what the actual outcomes are. Let's think about modalities at the society, political, market level that would facilitate individuals achieving their objectives more or less. And what could be a better description of love than 'I'm going to take—what you want is what I want for you, and I'm going to help you achieve that.' Economics is the loving social science, is my take on what economists do best."

Discussion Starters

Harry Anthony Patrinos, Emiliana Vegas, Rohan Carter-Rau provide a summary of the evidence in "An Analysis of COVID-19 Student Learning Loss" (May 2022, World Bank Policy Research Working Paper 10033, https://documents1.worldbank.org/curated/en/099720405042223104/pdf/IDU00f3f0ca808cde0497e0b88c01fa07f15bef0.pdf). "Our final database consists of 35 robust studies and reports documenting learning loss, representing data from 20 countries . . . Most studies (32) find evidence of learning loss. Of the 35 studies reporting learning loss, 27 reported findings in a comparable effect size format. . . . The average learning loss across these studies is 0.17 standard deviation—which equates to over half a school year of learning loss."

Cynthia R. Greenlee tells the story of how Samuel Rumph developed and marketed the Elberta peach starting in 1875 in "Reinventing the Peach, the Pimento, and Regional Identity" (*Issues in Science and Technology*, Summer 2022, https://issues.org/reinventing-peach-pimento-regional-identity-georgia-greenlee). "Just how Rumph begat this new peach is uncertain. It was succulent and bright yellow with red markings. Its pit came out easily, and its fruit matured early in the season. That timing and its firmness were boons, and the trees yielded their large, handsome fruit prolifically. As historian Thomas Okie wrote in his rigorous and compelling study of how the peach became a Georgia icon, Rumph had produced the 'industrial peach,' a reliable producer that was reasonably good to eat, relatively resistant to pests and diseases, amenable to growing in different climes and soil, and easily transportable. As a pioneer of what would eventually become agribusiness, Rumph considered the whole peach, from grafting to delivery, and intervened at various stages in the supply chain. First, he bred the peach that took the world by storm. Then, as a member of the Georgia State Horticultural Society's committee on packing and shipping peaches, Rumph devoted himself to studying how to send peaches around the country. . . . In an effort to make shipping a precise science rather than a gamble, Rumph created a slatted crate that could be stacked and wheeled, founding the Elberta Crate Company. His unpatented invention spawned industrywide imitation, and he went on to invent a refrigerated railway car—also unpatented—that was widely used by fruit growers thereafter. . . . Railroads were booming across the South, buoyed by ample northern investment. And peach growers' earnings—and nurserymen's active involvement in politics—determined where railroads would go and stop."

Robert Schultz and Anna Stansbury provide some facts about "Socioeconomic Diversity of Economics PhDs" (March 2022, Peterson Institute for International Economics, WP-22-4, https://www.piie.com/sites/default/files/documents/wp22-4.pdf). "In this paper, we use data from the National Science Foundation's Survey of Earned Doctorates (SED), an annual census of all individuals who receive a research doctorate from an accredited US institution in a given academic year, to examine the socioeconomic background of economics PhD recipients in the United States and compare it with that of PhD recipients in other disciplines. . . . Our analysis of the SED data shows that economics is even more unrepresentative by socioeconomic background than the average PhD field. Among US-born PhD recipients over 2010–18, 65 percent of economics PhD recipients had at least one parent with a graduate degree, compared with 50 percent across all PhD fields (and 29 percent for the population of US-born BA recipients over the same period). At the other end of the spectrum, only 14 percent of US-born economics PhD recipients in 2010–18 were first-generation college graduates, compared with 26 percent across all PhD fields (and 44 percent among all US-born BA recipients). This makes economics the least socioeconomically diverse of any major field for US-born PhD recipients. And its socioeconomic diversity appears to have worsened over time: while economics has consistently been less socioeconomically diverse than both the other social sciences and the biological and physical sciences, since 2000 it has also diverged from mathematics and computer science, the other two least socioeconomically diverse large PhD fields."

AMERICAN ECONOMIC ASSOCIATION

BEST PRACTICES FOR ECONOMISTS:

BUILDING A MORE DIVERSE, INCLUSIVE, AND PRODUCTIVE PROFESSION

A more diverse profession fosters a more vibrant discipline.

See practical suggestions and supporting research from the *AEA Task Force on Best Practices* regarding actions all economists can and should take. With intention, we CAN make change.

- **CONDUCTING RESEARCH**
- **SERVING AS COLLEAGUES**
- **WORKING WITH STUDENTS**
- **LEADING DEPARTMENTS AND WORKPLACES**

More information at
www.aeaweb.org/resources/bestpractices

Committee on the Status of Minority Groups in the Economics Profession

AMERICAN ECONOMIC ASSOCIATION

• Established 1968 •

SUPPORTING DIVERSITY IN ECONOMICS

The Committee on the Status of Minority Groups in the Economics Profession (CSMGEP) was established by the American Economic Association (AEA) in 1968 to increase the representation of minorities in the economics profession, primarily by broadening opportunities for the training of underrepresented minorities.

CSMGEP Programs

- Summer Economics Fellows Program
- Mentoring Program
- Summer Training Program
- Initiatives for Diversity and Inclusion

www.csmgep.org

SUPPORTING LGBTQ+ INDIVIDUALS IN THE ECONOMICS PROFESSION

The Committee on the Status of LGBTQ+ Individuals in the Economics Profession (CSQIEP) was established by the American Economic Association in 2019 (following three years of support of a related Ad Hoc Working Group on LGBTQ+ Economics) to address issues facing LGBTQ+ economists.

The Committee supports the AEA by offering recommendations on Best Practices concerning sexual orientation and gender identity issues in economics, supports LGBTQ+ economists through a variety of professional development and mentoring opportunities, and works in tandem with CSMGEP, CSWEP, and related AEA committees to advance equity, inclusion, and diversity throughout the economics profession.

CSQIEP Activities

- Mentoring Program
- Virtual Seminar Series
- Pink Papers at the AEA Annual Meeting
- Maynard's Notes LGBTQ+ Newsletter

@LGBTQ_Econ

www.csqiep.org

THE COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION

CSWEP

Advancing the Status of Women in the Economics Profession

- Publishes an annual survey of the representation of women in economics
- Offers mentoring workshops for junior faculty, graduate students, and early-career researchers outside of academia
- Conducts programs at the AEA and at meetings of other professional associations
- Co-sponsors the summer economics fellows program
- Hosts career development webinars
- Celebrates outstanding economists who have furthered the status of women in the profession
- Publishes *CSWEP News*. Free newsletter subscription available at **info@cswep.org**

www.CSWEP.org **@AEACSWEP**

Why an AEA Ombudsperson?

- Informal
- Confidential
- Impartial
- Independent

If you have experienced or witnessed incidents of harassment or discrimination, the AEA Ombudsperson can help guide you.

The AEA Ombudsperson enables AEA members to speak anonymously with a neutral third party about potential violations of AEA policies and determine if a formal or an informal report is appropriate. Whether the member has been personally impacted by misconduct or was witness to it, the ombudsperson is a confidential resource who can advise the member about additional resources and possible next steps. No action is ever taken without the complainant's authorization.

Need Assistance? Please Reach Out.

Leto Copeley, AEA Ombudsperson

aeaombuds@whiteandstradley.com

919-844-0400

https://copeleylaw.com/aea-ombudsperson-contact/

White & Stradley, PLLC
3105 Charles B Root Wynd
Raleigh, NC 27612

Formal Complaint?

Contact with the AEA Ombudsperson does not constitute official notice to the AEA or trigger a formal complaint. Formal complaints are filed through the AEA and are reviewed and investigated by the AEA Ethics Committee. The AEA Ombudsperson can provide guidance about the formal process and may separately advise the Association about observations of emerging patterns, repeat perpetrators, or systemic indicators.

Find more information about these policies and the AEA Ombudsperson at

www.aeaweb.org/ombudsperson

AMERICAN ECONOMIC ASSOCIATION

The AEA is dedicated to improving the climate of the economics profession by addressing harassment and discrimination, which are in violation of AEA policies and the AEA Code of Conduct.

AEA INITIATIVES FOR DIVERSITY AND INCLUSION

For more details and information regarding how to apply for AEA diversity initiatives, please visit **www.aeaweb.org/go/diversity-initiatives**

The American Economic Association is committed to the continued improvement of the professional climate in economics. In cooperation with key committees, the Association has launched several initiatives to support and promote diversity and inclusion in our profession.

1 AEA Award for Outstanding Achievement in Diversity and Inclusion

This annual award will recognize departments and organizations that demonstrate outstanding achievement in diversity and inclusion practices. Focus will be on those applicants that take productive steps to establish new programs and procedures to create an inclusive environment, and to increase the participation of underrepresented racial/ethnic minorities, women, and LGBTQ+ individuals.

2 Departmental Seed Grants for Innovation in Diversity and Inclusion

These grants, in amounts up to $5,000, will be awarded to economics departments to help establish new bridge programs or training programs for underrepresented minorities (URM). For example, a department might create a mentoring program for URM graduate or undergraduate students, create opportunities for URM students to do meaningful research assistant work, or start a program allowing URM students who need additional preparation for graduate school to take a lighter class load in the first year or to take core economics courses over two years.

3 The Andrew Brimmer Undergraduate Essay Prize

Thanks to the generosity of an anonymous donor, this paper prize has been established in honor of Andrew Brimmer, the first African American to serve on the Board of Governors of the Federal Reserve. The annual award will be presented to an undergraduate student at a US-based institution of higher learning majoring in economics, political science, public policy, or related fields for the best essay on the "economic well-being of Black Americans." The winner will receive a check for $1,000 and a plaque from the president of the AEA.

4 URM Travel Grants

This award is open to junior economics faculty members from traditionally underrepresented groups in the economics profession. The grants will advance career and professional development by defraying the costs of travel, lodging, and conference registration to attend the annual ASSA Meeting.

5 Small Group Breakfast Meeting for URM

Each year at the ASSA Meeting there will be a breakfast held with scholars from underrepresented minorities and prominent economists in attendance. The goal is to allow URM scholars access to AEA journal editors, executive board members, thought leaders in specific areas of economics, or other economists for the purpose of addressing issues of access to journals, conferences, and networks that are often out of reach for URM scholars.

6 Professional Development Grant for URM

This $2,000 grant was established to help advance the career and professional development of URM in the field of Economics. The award is open to eligible junior economics faculty members. Entrants to the essay competition should detail their research and how it relates to economics education.

These initiatives are another important step in helping make our field accessible and welcoming to anyone with the interest and ability to make a career in it. Please help us share this information throughout the profession so we can all work together and continue to improve.

AEA SUMMER TRAINING & SCHOLARSHIP PROGRAMS

HOWARD UNIVERSITY

May 25–July 23, 2023 • Washington, D.C.

INTENSIVE TWO-MONTH RESIDENTIAL PROGRAM

TWO LEVELS OF STUDY

COURSE WORK IN MATH, MICROECONOMICS, ECONOMETRICS, AND RESEARCH METHODS

REAL-WORLD EXPERIENCES UNIQUE TO THE NATION'S CAPITAL

PRESENTED IN COLLABORATION WITH:
Women's Institute for Science, Equity and Race, and the Federal Reserve Board as Economics Faculty Collaborators

SUBMISSION DEADLINE:
January 31, 2023

"My experience with the AEA Summer Program's rigorous curriculum and research project with the guidance of outstanding professors broadened my understanding of what it means to be a researcher and connected me with a powerful network of aspiring economists and economists of color."

Fanta Traore
HU (B.A. '15)
AEASP 2017

"It is inspiring to see the caliber of research the students are able to produce by the end of the program."

Jevay Grooms
AEASP 2016, 2017, and 2021
AEASP Research and Faculty Fellow

FOR MORE INFORMATION:
E-mail aeasp@howard.edu
http://economics.howard.edu/aeasp

Howard University and the American Economic Association extend a warm thanks to our 2022 sponsors.

National Economic Association 50th Anniversary
Association for Economic Research of Indigenous Peoples
PETER G. PETERSON FOUNDATION
NSF
The MIT Press
WISER Women's Institute for Science, Equity and Race
BATES WHITE Economic Consulting
Stata
amazon
ANALYSIS GROUP
ASHE American Society of Hispanic Economists

JOE NETWORK

The career network designed by economists—for economists.

The AEA's JOE Network is the preferred hiring tool for the economics job market, matching qualified candidates with relevant economics positions in leading institutions. Over 1,700 positions are filled each year from a pool of 5,100+ candidates.

Job Seekers

Showcase your profile to hundreds of hiring managers and quickly apply to your preferred positions. The JOE Network allows you to securely request reference letters and share job market materials.

Employers

Target search requirements to locate your preferred candidates, set up hiring committee access, and easily manage job postings and applications.

Faculty Letter-Writers

Reply to reference requests with default or customized letters, assign proxies, and easily monitor task completion. Over 150,000 requests are fulfilled on the JOE Network each year.

Use the JOE Network as your comprehensive tool for the economics job market!

AMERICAN ECONOMIC ASSOCIATION

aeaweb.org/JOE

The Journal of Economic Perspectives:

Proposal Guidelines

Considerations for proposing topics and papers for *JEP*

Articles appearing in the journal are primarily solicited by the editors and associate editors. However, we do look at all unsolicited material. Due to the volume of submissions received, proposals that do not meet *JEP*'s editorial criteria will receive only a brief reply. Proposals that appear to have *JEP* potential receive more detailed feedback. Historically, about 10–15 percent of the articles appearing in our pages originate as unsolicited proposals.

Readers are also welcome to send e-mails suggesting topics for *JEP* articles and symposia and to propose authors for these topics. If the proposed topic is a good fit for *JEP*, the *JEP* editors will work to solicit paper(s) and author(s).

Correspondence regarding possible future rticles for *JEP* may be sent (electronically please) to the assistant managing editor, Bradley Waldraff at b.waldraff@aeapubs.org. Papers and paper proposals should be sent as Word or pdf e-mail attachments.

Philosophy and style

The *Journal of Economic Perspectives* attempts to fill part of the gap between refereed economics research journals and the popular press, while falling considerably closer to the former than the latter. **The focus of JEP articles should be on understanding the central economic ideas of a question, what is fundamentally at issue, why the question is particularly important, what the latest advances are, and what facets remain to be examined.** In every case, articles should argue for the author's point of view, explain how recent theoretical or empirical work has affected that view, and lay out the points of departure from other views.

We hope that most *JEP* articles will offer a kind of intellectual arbitrage that will be useful for every economist. For many, the articles will present insights and issues from a specialty outside the readers' usual field of work. For specialists, the articles will lead to thoughts about the questions underlying their research, which directions have been most productive, and what the key questions are.

Articles in many other economics journals are addressed to the author's peers in a subspecialty; thus, they use tools and terminology of that specialty and presume that readers know the context and general direction of the inquiry. By contrast, **this journal is aimed at all economists, including those not conversant with recent work in the subspecialty of the author**. The goal is to have articles that can be read by 90 percent or more of the AEA membership, as opposed to articles that can only be mastered with abundant time and energy. Articles should be as complex as they need to be, but not more so. Moreover, the necessary complexity should be explained in terms appropriate to an audience presumed to have an understanding of economics generally, but not a specialized knowledge of the author's methods or previous work in this area.

The *Journal of Economic Perspectives* is intended to be scholarly without relying too heavily on mathematical notation or mathematical insights. In some cases, it will be appropriate for an author to offer a mathematical derivation of an economic relationship, but in most cases it will be more important that an author explain why a key formula makes sense and tie it to economic intuition, while leaving the actual derivation to another publication or to an appendix.

JEP does not publish book reviews or literature reviews. Highly mathematical papers, papers exploring issues specific to one non-US country (like the state of agriculture in Ukraine), and papers that address an economic subspecialty in a manner inaccessible to the general AEA membership are not appropriate for the *Journal of Economic Perspectives*. Our stock-in-trade is original, opinionated perspectives on economic topics that are grounded in frontier scholarship.

The *JEP* is freely available at **www.aeaweb.org/journals/jep**.

Guidelines for preparing *JEP* proposals

Almost all *JEP* articles begin life as a two- or three-page proposal crafted by the authors. If there is already an existing paper, that paper can be sent to us as a proposal for *JEP*. However, given the low chances that an unsolicited manuscript will be published in *JEP*, no one should write an unsolicited manuscript intended for the pages of *JEP*. **Indeed, we prefer to receive article proposals rather than completed manuscripts.** The following features of a proposal seek to make the initial review process as productive as possible while minimizing the time burden on prospective authors:

- Outlines should begin with a paragraph or two that precisely states the main thesis of the paper.
- After that overview, an explicit outline structure (I., II., III.) is appreciated.
- The outline should lay out the expository or factual components of the paper and indicate what evidence, models, historical examples, and so on will be used to support the main points of the paper. The more specific this information, the better.
- The outline should provide a conclusion.
- Figures or tables that support the article's main points are often extremely helpful.
- The specifics of fonts, formatting, margins, and so forth do not matter at the proposal stage. (This applies for outlines and unsolicited manuscripts).
- Sample proposals for (subsequently) published *JEP* articles are available on request.
- For proposals and manuscripts whose main purpose is to present an original empirical result, please see the specific guidelines for such papers below.

The proposal provides the editors and author(s) an opportunity to preview the substance and flow of the article. For proposals that appear promising, the editors provide feedback on the substance, focus, and style of the proposed article.

After the editors and author(s) have reached agreement on the shape of the article (which may take one or more iterations), the author(s) are given several months to submit a completed first draft by an agreed date. This draft will receive detailed comments from the editors as well as a full set of suggested edits from *JEP*'s Managing Editor. Articles may undergo more than one round of comment and revision prior to publication.

Guidelines for empirical papers submitted to *JEP*

JEP is not primarily an outlet for original, frontier empirical contributions; that's what refereed journals are for! Nevertheless, *JEP* occasionally publishes original empirical analyses that appear uniquely suited to the journal. In considering such proposals, the editors apply the following guidelines (in addition to considering the paper's overall suitability):

1. The paper's main topic and question must not already have found fertile soil in refereed journals. *JEP* can serve as a catalyst or incubator for the refereed literature, but it is not a competitor.
2. In addition to being intriguing, the empirical findings must suggest their own explanations. If the hallmark of a weak field journal paper is the juxtaposition of strong claims with weak evidence, a *JEP* paper presenting new empirical findings will combine strong evidence with weak claims. The empirical findings must be robust and thought provoking, but their interpretation should not be portrayed as the definitive word on their subject.
3. The empirical work must meet high standards of transparency. *JEP* strives to only feature new empirical results that are apparent from a scatter plot or a simple table of means. Although *JEP* papers can occasionally include regressions, the main empirical inferences should not be regression-dependent. Findings that are not almost immediately self-evident in tabular or graphic form probably belong in a conventional refereed journal rather than in *JEP*.

Join Us in **Shaping the Future of Economics**

Become an AEA Member!

As a member, your partnership with the AEA provides you with distinct career advantages and enables us to elevate the profession for all economists.

- Build lifelong relationships with mentors and peers
- Access essential research and job resources
- Publish and present your research to colleagues
- Gain meaningful support and advice on your work
- Collaborate on solutions to issues you care about

Today our community consists of 20,000+ individuals who are committed to the advancement of economics and its enduring contributions to society. We welcome the opportunity to help you advance your career in economics.

Aim High. Achieve More. Make a Difference.

AEA Journals @AEAJournals / @AEAInformation

AMERICAN ECONOMIC ASSOCIATION

Join today!
aeaweb.org/membership

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary access to JEP articles, go to the AEA website: http://www.aeaweb.org. Annual dues for regular membership are $24.00, $34.00, or $44.00, depending on income; for an additional fee, you can receive this journal, or any of the Association's journals, in print. Change of address notice must be received at least six weeks prior to the publication month.

The following Statement of Ownership, Management and Circulation is provided in accordance with the requirements, as contained in 39 U.S.C. 3658. *Journal of Economic Perspectives* is owned, managed, and published by the American Economic Association, a nonprofit educational organization, located at 2014 Broadway, Suite 305, Nashville, Davidson County, TN 37203. The Editor is Heidi Williams. The Managing Editor is Tim Taylor: *Journal of Economic Perspectives*, 2403 Sidney Street, Suite 260, Pittsburgh, PA 15203. The tax status of the American Economic Association has not changed during the preceding twelve months. During the preceding twelve months, the average number of copies printed for each issue was 3,266; the average total paid and/or requested circulation, 2,924; the average total non-requested distribution, 0; the average number of copies not distributed, 342; the average total distribution, 2,924. Corresponding figures for May 2022, the issue published nearest to filing date: total number of copies printed, 3,150; total paid and/or requested circulation, 2,763; total non-requested distribution, 0; number of copies not distributed, 342; total distribution, 2,763. During the preceding twelve months, the average number of requested and paid electronic copies each issue was 905; the total average requested and paid print and electronic copies, 3,829. Corresponding figures for May 2022, the issue published nearest to filing date: number of requested and paid electronic copies, 885; the total requested and paid print and electronic copies, 3,648. Certified by Barbara Fiser, Director of Finance and Administration.

The Journal of

Economic Perspectives

Fall 2022, Volume 36, Number 4

Symposia

Labor Market Institutions

The Size of Government Debt

Articles

Features

AMERICAN ECONOMIC ASSOCIATION