The Journal of Economic Perspectives

A journal of the
American Economic Association

Spring 2020

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.

No responsibility for the views expressed by the authors in this journal is assumed by the editors or by the American Economic Association.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Spring 2020, Vol. 34, No. 2. 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 34 • Number 2 • Spring 2020

Symposia

One Hundred Years of Women's Suffrage

Perspectives on Racial Discrimination

How Taxes Affect Location Choices

The Departure of Communism

Articles

Feature

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.

Votes for Women: An Economic Perspective on Women's Enfranchisement

Carolyn M. Moehling and Melissa A. Thomasson

On August 18, 1920, the United States granted women suffrage when the Tennessee House of Representatives voted to ratify the Nineteenth Amendment to the US Constitution—by a margin of one vote. The ratification of the "Susan B. Anthony" amendment marked the end of a nearly 80-year struggle on the part of women to gain the right to vote. Along the way, there had been some successes. The territory of Wyoming gave full suffrage to women in 1869, followed by the territory of Utah in 1870. By the time the US Congress passed the Nineteenth Amendment, 15 states had granted women full suffrage, with 13 of these in the West. Many more had given women partial suffrage, allowing them to vote in municipal or school elections and, in some cases, US presidential elections.

Economists, sociologists, political scientists, and others have long sought to explain the factors underlying the timing and the success of the women's suffrage movement. An extension of voting rights can fundamentally affect the distribution of resources in a society. Theories of suffrage extension seek to explain why groups in power would choose to share this power with the disenfranchised. All of these theories predict that men extend the franchise to women when the benefits of doing so outweigh the costs, but they differ in the benefits and costs they consider. In the next section, we discuss the history of women's suffrage in the United States, emphasizing the ways in which the movement for women's suffrage was intertwined

■ *Carolyn M. Moehling is Professor of Economics, Rutgers University, New Brunswick, New Jersey. Melissa A. Thomasson is Julian Lange Professor of Economics, Miami University, Oxford, Ohio. Both authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are cmoehling@economics.rutgers.edu and thomasma@miamioh.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.2.3.

with other historical episodes, movements, and shifting political and social landscapes. For example, we recount the rise of the women's rights movement and its relationship to the antislavery movement in the first half of the nineteenth century, and we discuss the impact of the decision to grant voting rights to black men, but not to women, after the US Civil War. We also examine the development of economic rights for women and explain why these rights evolved differently than voting rights. We describe the efforts of women's suffrage organizations to build support for the movement and to put pressure on state and federal legislators to enact suffrage legislation. Finally, we discuss the forces leading to the passage of the Nineteenth Amendment.

Throughout this narrative, we review the various theories that seek to explain why men in power chose to enfranchise women. Some of these models are based on competition among political elites and the desire to promote a particular policy agenda, others emphasize patterns of coalition-building, while still others stress how suffragists worked to create a greater demand for change. In the final section, we examine the empirical studies that use variation in the timing of suffrage across the states to distinguish between these theories. Perhaps unsurprisingly, no single factor or model can universally explain women's enfranchisement.

An Historical Overview of Women's Suffrage in the United States

Voting Rights in the Early Republic

The US Constitution as originally ratified in 1787 made no mention of gender and left the determination of voting rights to the states. However, the Constitution linked voter eligibility for the US House of Representatives to the qualifications required to vote for the "most numerous branch" of a state's legislature. Voting for presidential elections was even less well-defined: Article II, Section 1 specifies only that each state is allowed "to appoint, in such manner as the legislature thereof may direct, a number of electors."[1]

Early on, most states had statutes or constitutional provisions restricting suffrage to males, and states also tied the right to vote to property ownership or payment of taxes. Southern states further restricted suffrage to white males. Only New Jersey allowed women to vote; however, because married women were not allowed to own property, the state's property ownership restriction essentially limited voting to wealthy single women and widows. Nonetheless, women did participate in early elections, and political parties even recruited women voters to swing the vote in contested districts (Klinghoffer and Elkis 1992, 176).

In 1807, women in New Jersey lost the vote when the state legislature enacted a statute that restricted voting to white male tax-paying citizens. The political dynamic

[1] Under the original provisions of the Constitution, state legislatures selected members of the US Senate. This changed in 1913 with the ratification of the Seventeenth Amendment, which specified that senators be selected by popular vote.

within the Republican Party, divided between liberals in the North and moderates in the South, drove this change. Although the liberal Northerners had the upper hand, they knew they needed a united party to face the Federalists in the next presidential election. Moderate Republicans wanted to exclude non-taxpayers from voting, and growing nationalism within the party led to the exclusion of noncitizens. These two groups, however, traditionally voted Republican, so to balance the scales, single women and blacks, who traditionally voted for the Federalists, lost the franchise as well (Klinghoffer and Elkis 1992, 188).

New Jersey was not unique in narrowing voting rights in this period. Over the nineteenth century, states contracted as well as extended voting rights. These changes were politically motivated and reflected competition between parties and changing voter sentiments. As new states entered the Union without property ownership restrictions, some older states dropped such restrictions, too. However, states tended to establish or retain requirements that voters must be taxpayers (McConnaughy 2013, 21), and many states enacted restrictions that took voting rights away from racial minorities, immigrants, and individuals who were illiterate or had criminal histories. [2]

Abolition, Seneca Falls, and the Women's Rights Movement before the Civil War

The women's rights movement grew out of the broader reform movements of the 1830s. As women activists fought for the rights of others, they began to recognize and elucidate the constraints placed on them because of their sex (DuBois 1987, 837). Female abolitionists, in particular, were criticized for intruding on men's affairs and not following social conventions. As the antislavery movement debated whether women should be allowed leadership positions, women activists began to articulate some parallels between their situation and that of enslaved persons. In an 1837 piece published in the weekly abolitionist newspaper, *The Liberator*, Sarah Grimké wrote that women also experienced the "irrepressible desire for mental and spiritual freedom which glows in the breast of many who hardly dare speak the sentiments…" (as quoted in Buhle and Buhle 2005, 6). Such statements were met with rebukes from men and women alike. Catharine Beecher, sister of Harriet Beecher Stowe, sharply criticized "those who are bewailing themselves over the fancied wrongs and injuries of women in this nation." She argued that women and men occupied "separate spheres," and women's role was to care for the home and children (DuBois 1987, 838). By the end of the 1840s, many women abolitionists recognized that they needed to use the political system to secure equal rights for women as well.

In July 1848, Elizabeth Cady Stanton and Lucretia Mott called a convention in Seneca Falls, New York "to discuss the social, civil, and religious condition and rights of woman." Almost 300 people attended the convention, many of whom were men (Keyssar 2009, 240). Stanton drafted the Declaration of Sentiments to summarize

[2] For a history of voting rights in the United States, see Keyssar (2009).

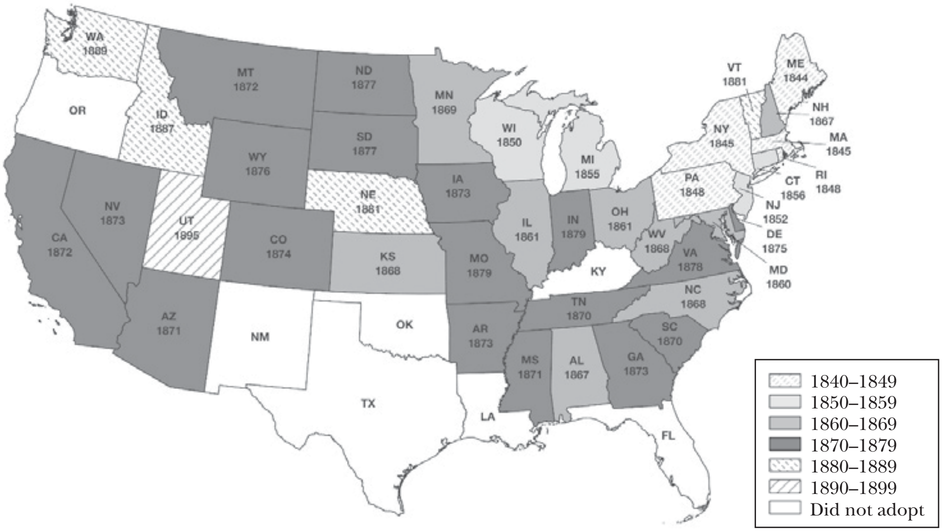
the resolutions coming out of the convention, framing it to mirror the Declaration of Independence. The Declaration of Sentiments asserts that "all men and women are created equal" and then goes on to enumerate the "repeated injuries and usurpations on the part of man toward woman," with the first grievance listed as "He has never permitted her to exercise her inalienable right to the elective franchise," followed by "he has compelled her to submit to laws, in the formation of which she had no voice." The list of further complaints includes the restrictions on women's economic activities and the imbalance of power in marriage (Buhle and Buhle 2005, 95–96).

Of the rights demanded for women, suffrage was the most radical. In their *History of Woman Suffrage*, Anthony and Stanton noted that the resolution encouraging women to fight for suffrage was the only one *not* unanimously adopted at Seneca Falls. Some convention-goers feared that including suffrage would lead to more opposition to their other demands for equality in marriage and economic pursuits "and make the whole movement ridiculous" (Buhle and Buhle 2005, 97). Stanton and the prominent abolitionist, Frederick Douglass, won the day by countering that "the power to choose ruler and make law, was the right by which all others could be secured" (Buhle and Buhle 2005, 97).

While the Seneca Falls Convention is often identified as the start of the women's suffrage movement, there were actually many meetings organized during the late 1840s and early 1850s to promote the rights of women. Seneca Falls looms large in popular memory in part because Elizabeth Cady Stanton and Susan B. Anthony wrote the first history of the movement (Keyssar 2009, 142), in which they asserted the primacy of the Seneca Falls convention and preserved a record of its proceedings for future generations.[3] By the early 1850s, suffrage had become the centerpiece of the women's rights agenda. At the Second National Woman's Rights Convention held in Worcester, Massachusetts in 1851, the first resolution made explicit why suffrage was the primary goal: "resolved, that while we would not undervalue other methods, the Right of Suffrage for Women is, in our opinion, the cornerstone of this enterprise, since we do not seek to protect woman, but rather to place her in a position to protect herself" (Buhle and Buhle 2005, 112). The justification for women's suffrage was based on the republican notion of equal rights for all. Activists argued that women were entitled to the same political rights as men; women,

[3] In 1876, Stanton and Anthony decided to write a history of the movement. They planned to spend four months and produce a pamphlet of a few hundred pages. By 1886, they had produced three volumes, each about 1,000 pages under the title *History of Woman Suffrage*. These volumes contained a mix of written accounts of events and primary documents like speeches, letters, newspaper articles, and reports. Anthony worked with Ida Husted Harper to edit a fourth volume published in 1902 that documented the history from 1883 to 1900. Harper edited two more volumes, both published in 1922, that documented the movement from 1900 to 1920. As Buhle and Buhle (2005, xix) argue, the *History of Woman Suffrage* "presents a defensive and highly partisan portrait of their own National Woman Association (NWSA)." They gave little space to their rival organization, the American Woman Suffrage Association, and other reform organizations involved in the women's suffrage cause.

Figure 1
Married Women's Property Laws Enacted Prior to 1900



Source: Khan (1996, Table 1, 363-64).
Note: The data displayed refer to the enactment of laws granting married women control over property. Khan (1996) also provides the dates of laws giving married women control of their earnings and the right to engage in contracts and business without husband's consent.

like men, were subject to the decisions made by the government, so they too should have a direct voice in who governed (DuBois 1987, 841).

The Advancement of Economic Rights for Women

Despite the focus on gaining the vote, the early legislative successes of the women's movement were in eliminating coverture and thus securing the rights of married women to own property, to control their earnings, and to enter into contracts. Figure 1 presents data on the timing of laws that granted married women control over property. In most states, these property laws preceded, or were enacted in conjunction with, laws granting married women control over their earnings and laws allowing women to engage in contracts or business without their husbands' consent, often referred to as "sole trader" laws (Khan 1996, 362–64). Maine, Massachusetts, New York, Pennsylvania, and Rhode Island granted married women property rights in the 1840s, and other states in the Northeast and the Midwest followed in the 1850s and early 1860s. Most of the states in the West and South enacted these laws in the 1870s. The successful reforms in economic rights occurred in part because the increased focus on political equality broadened, rather than narrowed, the scope of the women's movement to include these areas. DuBois (1987, 842–43) notes that the idea that women deserved greater political voice naturally connected to the idea that women should have greater independence in other spheres, particularly in marriage.

Changes to property rights altered the relationship between men and women within marriage. Therefore, these changes must be discussed in terms of how the allocation of property rights within marriage affects household decisions. One theoretical approach asserts that women place a higher value on children's welfare than men, and men recognize that giving women more economic power will increase investments in children's education (Doepke and Tertilt 2009). When the returns to human capital are low, as they were in the preindustrial, mostly agricultural economy in the early nineteenth century, the benefits to men of being able to control their wives' property outweigh any benefits from investing more in children's education. As the returns to human capital increase, the scale tips toward more education for children, and men are more likely to grant women greater economic power.

An alternative, yet complementary, approach emphasizes how coverture reduces married women's incentives to pursue economic opportunities, and hence, reduces household earnings and wealth (Geddes and Lueck 2002). The costs to households of these disincentives grow as the economy develops and women have more opportunities in the formal labor market. Eventually, these costs lead men to grant women property rights. An empirical examination of the timing of the enactment of married women's property laws across states finds that the early movers were states with larger urban populations, higher female school enrollment, and higher wealth per capita, results that are consistent with both models (Geddes and Lueck 2002). Women's economic opportunities and the returns to human capital would have been greater in urban than rural areas, and the higher rates of female school enrollment indicate increased investments in expectation of these returns. Further supporting these theories, the enactment of married women's property laws led to increases in women's patenting of inventions. Once women had the legal right to control their property and earnings in marriage, they were more willing to engage in commercial activities (Khan 1996).

What is striking about the geographic diffusion displayed in Figure 1 is how different it is from that of women's suffrage. As we will discuss below, the early movers in granting married women property rights were slow to grant women the vote; many early movers did not grant women access to the ballot until the Nineteenth Amendment, demonstrating that the politics of these two dimensions of women's empowerment were quite different. While changes to property rights altered the relationship between men and women within marriage, they did not change political institutions directly.

Reconstruction and the Defeat of Universal Suffrage

The aftermath of the Civil War generated a broad discussion of political rights. Women's groups played a key role in organizing popular support for the Thirteenth Amendment, which abolished slavery. As the discussion turned to constitutional remedies to ensure the rights of the former enslaved population, women's rights advocates believed that such remedies should also extend to them. They argued that women, like the former enslaved, deserved the vote not because of their special status, but because they were human beings endowed with natural rights. As one activist put it, race and sex were just "two accidents of the body" unworthy of constitutional recognition (DuBois 1987, 845–46).

However, even radicals and abolitionists who had previously advocated for universal suffrage retreated when it became clear that opposition to women's enfranchisement could lead to the defeat of suffrage for black men (Keyssar 2009, 144). The Republican Party leadership sought to separate black men's suffrage from women's enfranchisement and prioritized gaining voting rights for black men. Prior to the ratification of the Fourteenth Amendment, abolitionists Wendell Phillips and Theodore Tilton suggested they focus on enfranchising blacks first and women later. At this, Susan B. Anthony became angry and declared that "she would sooner cut off her right hand than ask for the ballot for the black man and not for woman" (as quoted in Harper 2005, 261). Anthony and other prominent suffragists such as Elizabeth Cady Stanton vocally denounced the wording of the Fourteenth Amendment. In contrast, Frederick Douglass, a supporter of women's rights since the convention in Seneca Falls, famously stated in a debate at the American Equal Rights Association meeting in 1869 that when women "are dragged from their houses and hung upon lamp-posts…then they will have an urgency to obtain the ballot equal to our own" (Buhle and Buhle 2005, 258).

Republicans also anticipated political gains from giving black men the vote: it would give them a solid base of support in the South, as well as help their political strength in the North. McConnaughy (2013, 34–37) calls this "strategic enfranchisement"—when political actors seek new supporters from the disenfranchised. As Massachusetts Senator Charles Sumner asked his fellow Republicans (as quoted in Keyssar 2009, 74):

> You need votes in Connecticut, do you not? There are three thousand fellow-citizens in that state ready at the call of Congress to take their place at the ballot box. You need them also in Pennsylvania, do you not? There are at least fifteen thousand in that great state waiting for your summons…be assured they will all vote for those who stand by them in the assertion of Equal Rights.

In contrast, giving women access to the ballot was not expected to yield the same returns. Politicians viewed women as encompassing too much variation for their votes to be viewed as a bloc; many believed they would vote as their husbands, so the vote would just be doubled (McConnaughy 2013, 252).

The ratification of the Fourteenth Amendment dealt another serious blow to the movement to advance women's rights by specifying that the basis for a state's congressional representation was its number of male citizens. This was the first constitutional provision to discriminate explicitly based on sex, and it undermined women's claims to suffrage. The Fifteenth Amendment, ratified in 1870, delivered a second setback when it prohibited limiting a citizen's right to vote "on account of race, color, or previous condition of servitude," but notably *not* on account of sex.

Changing Directions for the Women's Suffrage Movement

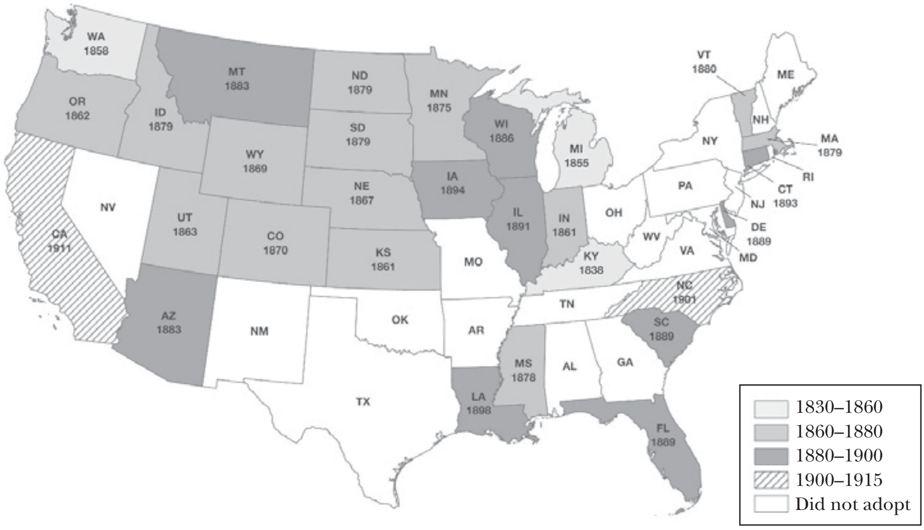
The defeat of universal suffrage during Reconstruction led to profound changes in the women's suffrage movement. Disagreement over strategy split the leadership of the movement into two separate organizations in 1869: the National Woman's Suffrage Association (NWSA) led by Elizabeth Cady Stanton and Susan B. Anthony, and the American Woman's Suffrage Association (AWSA) led by Lucy Stone and her husband, Henry Blackwell. Stanton and Anthony, embittered by the ratification of the Fourteenth Amendment, sought a federal amendment and focused on lobbying Congress. In contrast, Stone and Blackwell had supported the passage of the Fourteenth Amendment. Although disappointed by the exclusion of women, Stone noted: "I will be thankful in my soul if *any* body can get out of the terrible pit" (Stanton et al. 1881, 2 and 384). She and Blackwell believed the most productive strategy would be to work state by state to secure women the vote (Keyssar 2009, 149).

The arguments used to advance women's suffrage also began to shift. Although advocates continued to argue for women's enfranchisement in terms of women deserving the same rights as men, some began to argue women should be given the vote because of the ways in which they differed from men. These arguments built on the theory of "separate spheres." Previously, separate spheres had been used to oppose the expansion of rights for women by claiming that their focus on matters of the home made them unfit for political life. But suffrage supporters began to repurpose these arguments to claim instead that women's distinct experiences and perspectives would bring morality and virtue to politics and promote reform (DuBois 1987, 848–49; McCammon, Hewitt, and Smith 2004).

A few arguments for women's suffrage also appealed to the deeply seated racism of American society; proponents for women's suffrage criticized the Fifteenth Amendment for giving the vote to the "lowest classes of manhood" over "the higher classes of women" (DuBois 1987, 850). Stanton, Anthony, and other leaders of the women's movement are on record making explicitly racist arguments for women's suffrage. Some supporters of women's suffrage in the South argued that giving women the vote would counter the influence of black male voters (Buhle and Buhle 2005, xxv). Others drew on racist arguments to oppose women's suffrage, arguing that giving black women the vote would pose a threat to white supremacy (McConnaughy 2013, 171–72).

Led by Stanton and Cady, the National Woman's Suffrage Association tried to leverage the reform sentiment that generated the Fourteenth and Fifteenth Amendments to secure a federal amendment granting women the vote. In 1869, radical Republican George Julian of Indiana proposed a constitutional amendment linking the right to vote to citizenship. This provision would have taken away the power of states to define voting rights, and hence, it met with strong opposition. Susan B. Anthony drafted an amendment with a narrower scope: prohibiting discrimination in voting rights "on account of sex." In 1878, Senator Aaron A. Sargent of California introduced this amendment to Congress, and in 1882, committees from both houses recommended passage of the amendment. In 1887, the proposal was

Figure 2
Women's School Suffrage Prior to Nineteenth Amendment



Source: Nicholas (2018, Table 1, 463–67).

brought forward for a floor vote in the Senate, but the outcome devastated women's suffrage supporters: 16 yeas, 34 nays, and 26 abstentions. In every subsequent session of Congress, supporters reintroduced the amendment, but it did not again come to a floor vote until 1914 (Keyssar 2009, 150).

Partial Suffrage

The women's suffrage movement did make some gains at the state level in the 1800s; many states granted voting rights to women in some, but not all elections. For instance, as early as 1838, Kentucky legislators granted widows and single women with property the right to vote in elections involving school taxes. As shown in Figure 2, many states granted women partial suffrage by giving them the right to vote in school elections. The real push for school suffrage came in the decades after the Civil War. Legislators justified this type of partial extension of suffrage using a version of the separate spheres argument: women, as mothers, should be able to vote in elections that affected the welfare of children (Keyssar 2009, 150).

The extension of school suffrage to women, however, was also a legislative tactic to promote particular policy agendas for public education systems. In most states, changes in school suffrage were enacted as part of legislation defining the public provision of education (Nicholas 2018, 461–68). This connection between bringing in new voters and enacting new policy fits well with economic models that explain the expansion of voting rights as the outcome of political competition between the groups in power (Llavador and Oxoby 2005; Lizzeri and Persico 2004). These models suggest that if women have different policy preferences than men—or at least a different distribution of policy preferences than men—then political actors may extend the vote to women in order to promote their policy agenda. State political leaders believed that allowing women to vote in school elections would help to advance their goals for public education.

In addition to limiting women's political influence to particular policy issues, school suffrage and other forms of partial suffrage, such as giving women rights to vote in municipal elections, preceded the full enfranchisement of women in part because of the greater legislative costs associated with broader voting rights. Partial suffrage was procedurally easier to enact than changes to voting laws that affected offices named in state constitutions (Keyssar 2009, 150). Only in Delaware could a simple legislative vote change full voting rights. All other states required at least a referendum and others even more action. For example, in Illinois, full voting rights could only be changed with a positive legislative vote, followed by a favorable vote at a constitutional convention (which occurred only every 20 years) and a subsequent referendum.

However, US territories could enact suffrage with only a single legislative vote. This relative procedural ease may explain why a number of territories fully enfranchised women before states: when pro-suffrage groups managed to bring bills to the floor, they had a higher chance of being enacted (McCammon and Campbell 2001, 65). The territory of Wyoming gave women full suffrage in 1869, followed by the territory of Utah in 1870. Despite these early victories, the push for full suffrage stalled. In the 1880s, only the territories of Washington and Montana extended women full suffrage. But then, Utah repealed suffrage under the Edmunds-Tucker Act in 1887 and did not restore it until 1895. Washington enacted and revoked suffrage several times from 1887 to 1910.[4]

These limited and sometimes temporary victories mask the enormous efforts made by pro-suffrage organizations during this period. Pro-suffrage groups launched numerous campaigns, and a number of states held referenda in the 1870s and 1880s. Although most of these initiatives met with defeat, many men voted for them. And as Keyssar (2009, 151) notes, in states where the question of women's suffrage did not come to a referendum, "suffrage organizations were active, state legislators were obligated to vote on suffrage bills year after year, and support… often cut across party lines."

Increased Efforts for State-Level Change, 1890–1912

In 1890, frustrated by the lack of progress on broader suffrage for women, the National Woman's Suffrage Association and the American Woman's Suffrage

[4] In 1887, the Washington Territorial Supreme Court revoked suffrage for women. The legislature then passed a new law granting women suffrage in 1888, which became nullified when the court ruled on a second suffrage suit later that same year. After achieving statehood in 1889, supporters worked to amend the Washington Constitution to give women the right to vote in 1898 and 1906 before finally succeeding in 1910.

Association merged to form the National American Woman Suffrage Association (NAWSA). The new organization stepped up the pressure on the states to enact women's suffrage. Between 1890 and 1920, the NAWSA engaged in hundreds of campaigns to promote women's suffrage (McDonagh and Price 1985, 416).

Women's suffrage organizations pursued the usual channels to effect policy change; they set up political lobbies and campaigned for political candidates who supported women's suffrage. These efforts forced state legislatures to take notice; in the states in which the organizations engaged in these political activities, more suffrage bills were introduced in legislative sessions and more of these bills were put to roll call votes, even if they were never passed into law (King, Cornwall, and Dahlin 2005).

Women's suffrage organizations also worked to broaden support for their cause by forming coalitions with other social movements. These coalitions generated more support for women's suffrage not only among women but also among men, who could use their votes to make policymakers pay attention. Unlike the post-Civil War extension of suffrage to black males that promised to add new voters to the Republican Party, the extension of the vote to women did not offer clear benefits to either of the major political parties. By forming alliances with other social movements, women's suffrage organizations were able to create greater electoral pressure for enfranchising women. This strategy was particularly effective when political contests were close and when a third party threatened to upset the balance of power between the two major parties (McConnaughy 2013, 34–37). Of course, these alliances could also generate more opposition as well.

The women's suffrage movement found powerful allies in labor and farm organizations. In 1870, just under 13 million women, or 14.8 percent of females aged 10 or older, were in the labor force. By 1900, these numbers had increased to 29 million women, or 18.3 percent of females aged 10 or older (Hooks 1947, 34). As their workforce participation increased, women became increasingly involved in the labor movement. Some independent unions admitted women before 1870, and the Knights of Labor allowed women to be members beginning in 1881. By the 1890s, the larger labor unions began to endorse women's suffrage (McConnaughy 2013, 140). The Women's Trade Union League was founded in 1903, and the first significant strike of women workers occurred in 1909–1910 among the shirtwaist makers in New York and Philadelphia (Flexner and Fitzpatrick 1996, 234).

The alliances of women's suffrage organizations with labor and farming interests played particularly important roles in some state battles. For instance, in Illinois, the Farmers' Alliance paved the way for women's access to school ballots in 1891—the first successful suffrage reform in the state. As described by McConnaughy (2013, 146–48), the Alliance won three seats in the Illinois General Assembly, which gave them the legislative leverage to swing votes for or against the two major parties. This political power was key to winning women the right to vote in school elections.

The women's suffrage movement also found allies in the prohibition movement. The connection between women's suffrage and temperance was long-standing and preceded the movement's link to prohibition. A lack of economic rights placed married women at the mercy of their husbands. If husbands were heavy drinkers, their wives and children could be reduced to destitution with no redress. Many suffrage leaders were temperance workers before they became suffragists, including Susan B. Anthony, Elizabeth Cady Stanton, Lucretia Mott, and Lucy Stone. The leadership between the two movements often overlapped, although not all suffragists and supporters were in favor of the temperance and prohibition movements (Flexner and Fitzpatrick 1996, 174; McDonagh and Price 1985, 432).

Many of the women in the prohibition movement were conservative and embraced women's traditional roles in the household and society. However, over time, these women came to realize that they could not effect change without the vote. In this way, they also were drawn to the separate spheres argument for women's suffrage. This argument, rather than challenging women's traditional roles, made those roles the justification for giving women the vote (McCammon and Campbell 2002, 232).

An important conduit for the alliance of the two movements was the Woman's Christian Temperance Union (WCTU). The WCTU spearheaded the fight for prohibition and had extensive reach; by the end of the nineteenth century, it had branch organizations in every state and was the largest women's organization in the United States (McCammon and Campbell 2002, 232). Between 1874 and 1919, the WCTU formed coalitions with suffrage organizations in all but four states. In many instances, these coalitions formed shortly after the defeat of a prohibition measure, and in almost all cases, the collaboration took the form of the WCTU mobilizing its resources to promote women's suffrage rather than the women's suffrage organization working to promote prohibition. The women of the WCTU recognized that to achieve their objectives, they needed to have the vote (McCammon and Campbell 2002).

The link to the prohibition movement brought greater attention and support to the women's suffrage movement, but it proved to be a liability at times. A case in point is the defeat of a referendum on women's suffrage in California in 1896. Both the Populist and Republican parties supported the referendum, but ten days before the vote, representatives from the Liquor Dealer's League met in San Francisco and "resolved 'to take such steps as were necessary to protect their interests'" (Flexner and Fitzpatrick 1996, 216). The League sent letters to barkeepers, hotel owners, grocery proprietors, and others, urging them to vote against the amendment. When the vote was tallied, women's suffrage was carried in all counties except San Francisco and Alameda, where the opposition was strong enough to defeat the amendment.

On balance, it is not clear whether the alliance with prohibition interests was a net benefit or liability for the women's suffrage movement. The alliance clearly motivated "wet" interests to mobilize their resources to fight women's suffrage. However, it also expanded the base of support for women's suffrage, and expanding the base of support was key to propelling the movement forward.

After a flurry of successes in the early 1890s, the legislative progress of the suffrage movement had stalled, leading to a period often referred to as the "the doldrums." Suffrage leaders adopted new strategies to strengthen the movement's organizational structure and extend its outreach. They followed the model of successful political machines, setting up operations in towns across the country and at the ward level in major cities and going door-to-door to distribute pamphlets and broadcast their message (Keyssar 2009, 162). Their most attention-grabbing tactics involved taking the movement's message to the streets—literally. One tactic was "street speaking," in which suffrage advocates stood on soapboxes on street corners or on the backs of automobiles and argued their case to whomever was walking by. Suffrage parades proved even more effective, though, for getting the suffrage message to the broader public. Suffragists, all dressed in white, marched in formation carrying banners and signs calling for "Votes for Women!" In 1908, between 200 and 300 women marched in the first suffrage parade in Oakland, California. The women marched down the streets to the convention of the California Republican Party to demand support for women's suffrage. Other suffrage organizations built from this model, staging even larger suffrage parades. The suffrage parade in New York City in 1915 involved an estimated 20,000 to 25,000 women, including 74 women on horseback, 57 marching bands, and 145 decorated automobiles (McCammon 2003, 791). These parades brought together suffrage supporters from across the political and economic spectrum and put this diversity on display to the public. The scale and spectacle of the parades led to coverage by the press, greatly expanding public awareness of the movement.

The doldrums of the suffrage movement ended in 1910 when the state of Washington granted women full suffrage. California followed in 1911, and Arizona, Kansas, and Oregon followed in 1912. The victory in California reflected in part the improved organizational structure of the movement. Suffrage supporters focused on organizing in small towns and rural areas, where they knew they had stronger support. They also hired detectives and guards to prevent liquor interests from sabotaging ballot boxes. While they were strongly defeated in urban areas, the referendum won by 3,587 votes, "an average majority of one vote in every voting precinct in the state" (Flexner and Fitzpatrick 1996, 249).

The Move to a Federal Amendment

With the exception of New Mexico, all of the states west of the Rocky Mountains extended full voting rights to women by 1914. In contrast, east of the Rockies, only Kansas enacted full women's suffrage before 1914. Yet the national political landscape began to change in 1912 when divisions in the Republican Party led to the birth of the Progressive "Bull Moose" Party, led by the former president, Theodore Roosevelt. Endorsing women's suffrage fit well with the reform platform of the Progressive Party and also helped the party appeal to labor interests. Roosevelt even presented the party's case for women's suffrage in terms of the need of workingwomen to have the ballot just like workingmen. By supporting women's suffrage, the Progressive Party was also able to tap into the organizational structure of the suffrage movement as it hurriedly staged its campaign for the November 1912 presidential election (McConnaughy 2013, 238–39).

Roosevelt won 27 percent of the vote in that election, forcing the two major parties to take note and creating an opening for a renewed drive for a federal women's suffrage amendment. Progressive Party candidates also won seats in state legislatures, giving the women's suffrage movement renewed leverage for state-level legislative change. In Illinois, for example, the Progressives won 26 seats in the state House and two in the state Senate. The defection from the Republican Party had also led to the Democratic candidate winning the gubernatorial race in Illinois (McConnaughy 2013, 157). The challenges of pursuing an amendment to the state constitution led the Progressives to develop an ingenious plan: granting women presidential suffrage. The US Constitution only specified that states were "to appoint, in such manner as the legislature thereof may direct" electors for presidential elections. This meant that unlike voting for the US Congress, there were no links between who could vote in presidential elections and state-level elections. Presidential suffrage, therefore, could be granted by an act of the state legislature alone (McDonagh and Price 1985, 417). In 1913, Illinois became the first state to grant women presidential suffrage. Sixteen states followed before 1920, including Texas and Tennessee (Keyssar 2009, 367).

The federal amendment came to a floor vote in the Senate in 1914 and the House of Representatives in 1915. Although the tallies of both votes went against the amendment, support did not split along party lines. In both the Senate and the House, legislators from states where women had the vote were much more likely to support the amendment than legislators from other states (Jones 1991; McConnaughy 2013, 243). In the House, support was also positively correlated with the share of the Progressive Party in the 1912 presidential election and the ratio of men to women in a legislator's state and negatively correlated with the size of the liquor industry (Jones 1991, 430).

World War I created another shift in the political landscape. Under the leadership of Carrie Chapman Catt, the National American Woman Suffrage Association suspended its lobbying for women's suffrage in order to support the war effort. However, a splinter group, the National Woman's Party, led by Alice Paul, intensified its campaign for a federal amendment by picketing the White House. Paul and many of her fellow Woman's Party activists were arrested and imprisoned. They responded by conducting hunger strikes and were force-fed, much to the shock of the American public. The differing approaches during the war of the two major wings of the women's suffrage movement served to keep women's suffrage in the public attention, while also reinforcing the ways women contributed to the nation's defense (Keyssar 2009, 172–73).

The roles women played during World War I served to shift sentiment for women's suffrage in their favor. As the House of Representatives debated the resolution that would ultimately become the Nineteenth Amendment, Congressman John MacCrate (R, NY) stated (58 Cong. Rec. 84, May 21, 1919):

> …whether you consider the franchise a right or a privilege, the women of America deserve the right, or they have earned the privilege. Everywhere you

> went during the past two years you saw women in uniform…in the Salvation Army, the Red Cross, the Knights of Columbus, the Young Men's Christian Association, Young Men's Hebrew Association, and other allied war activities… I submit to your judgement that the women of America have been as potential soldiers during the past war as have been the men of America.

The fact that many other countries had already granted suffrage to women may have also played a role in convincing lawmakers to vote for the Nineteenth Amendment. On the eve of the US vote in 1918 and 1919, several European countries, including Austria, Germany, Belgium, the Netherlands, and Sweden enfranchised women, while many others had granted women suffrage much earlier (Bertocchi 2011).[5] Just prior to the vote in which the Nineteenth Amendment passed the House of Representatives Congressman John Raker (D, CA) asked his colleagues (58 Cong. Rec. 82, May 21, 1919):

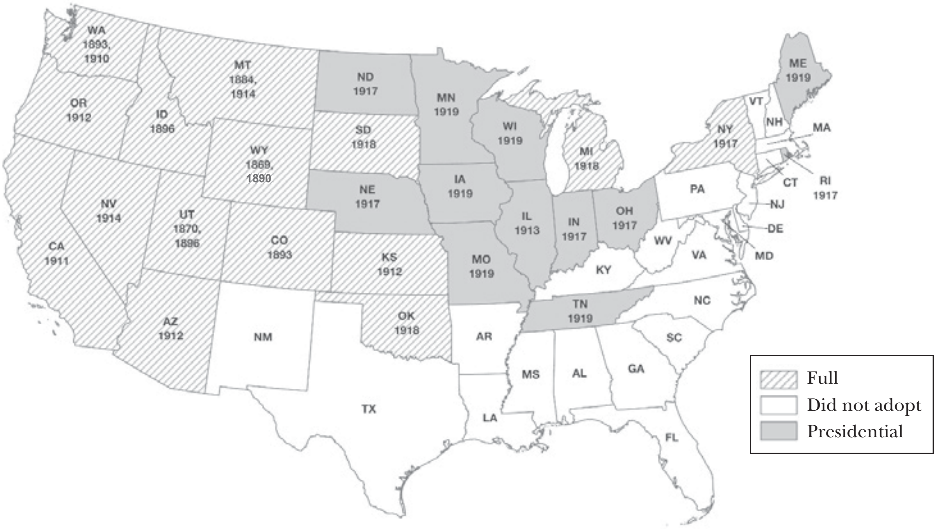
> …should we be the last of the civilized countries of the world to extend this right—we who boast that we stand for giving men the opportunity to express their voice in our Government, that we might have a Government of the people, not by heredity, but that the people might express their will and desire as to what their Government should be? Is it right that we should be the last?

Once again, legislators from states with full or presidential women's suffrage were more likely to support the amendment in the House, which passed the amendment on May 21, 1919. Given the success of the state-level legislative initiatives between 1915 and 1919, this amounted to many more votes in 1919 than it had in 1915. Support was also still higher from states with a greater ratio of men to women. But with the ratification of the Eighteenth Amendment prohibiting the sale and consumption of alcohol, the size of the liquor industry no longer had an effect on voting patterns. In the Senate, which passed the measure on June 4, 1919, the key predictors of votes were the share of Progressive Party voters in 1912 and the ratio of men to women in the state (Jones 1991, 430–33).

To take effect, 36 states had to ratify the Nineteenth Amendment. The last state to ratify the amendment was Tennessee. Despite Tennessee having just enacted a presidential suffrage law in 1919, the fight for ratification was fierce. Weiss (2018) provides a very engaging account of the battle for ratification in Tennessee. Some state legislators who had voted in favor of presidential suffrage in 1919 opposed the amendment, claiming it violated states' rights. In the end, the Tennessee legislature voted to ratify the amendment by a single vote on August 18, 1920.

[5] New Zealand granted full suffrage to women in 1893, followed by Australia, Finland, Norway, Denmark, and Canada. England granted suffrage to select women over the age of 30 in 1918 but did not give all women suffrage until 1930 (Bertocchi 2011).

Figure 3
Women's Full and Presidential Suffrage Prior to the Nineteenth Amendment



Source: Keyssar (2009, Table A.20, 368); McDonagh and Price (1985, Table 1, 417).

Empirical Analyses of the Diffusion of Women's Suffrage

As the narrative above highlights, many factors were at play in the struggle to extend the vote to women. Some of these factors, like competition between political parties and the opposition of liquor interests, could be classified as "supply side" in that they influenced the willingness of male legislators and male voters to grant women suffrage. On the "demand side" were the actions taken by women's suffrage organizations to generate support for their cause. A number of studies have leveraged the geographic variation in the timing of women's suffrage to identify empirically the relative importance of these different factors. Figure 3 presents a map showing the dates that women were granted full or presidential suffrage prior to the ratification of the Nineteenth Amendment. The Western states were the leaders in giving women the vote. All of the states that had granted full suffrage to women before World War I were east of the Mississippi River. This geographic pattern is very different than that displayed in the map in Figure 1 that shows the timing of laws granting married women property rights. Many of the urban, industrialized states of the Northeast that had moved early to extend women's economic rights did not extend voting rights to women prior to the Nineteenth Amendment.

Scholars have long debated why the Western states moved first to give women full suffrage. In 1967, the historian Alan Grimes argued that Western politicians, believing women could counter "frontier rowdiness," granted them access to the ballot in order to promote a respect for law and order and impose the "Puritan ethic" as the norm for community behavior (Grimes 1967, 76–77). Grimes's hypothesis anticipated the economic models described above that attribute the expansion of suffrage to the desire of one group in the political elite to promote a particular policy agenda (Llavador and Oxoby 2005; Lizzeri and Persico 2004).

McCammon and Campbell (2001) propose that it was the strength and strategies of women's suffrage organizations that led to the leadership of the West in women's suffrage. In an event-history analysis of the timing of full suffrage, they find no relationship between suffrage and the number of saloonkeepers per capita, which they interpret as evidence against Grimes' Puritan ethic hypothesis. However, the number of saloonkeepers per capita could also be interpreted as a measure of the opposition to prohibition, which as discussed above, was strongly linked to women's suffrage. McCammon and Campbell do find that suffrage passed earlier in states where suffrage organizations presented what they call "expediency" arguments for women's suffrage. These arguments, built on the philosophy of separate spheres, claimed that women would bring "special 'womanly' skills to politics to address public issues related to morality" (69). This finding does not refute the Puritan ethic hypothesis; rather, it suggests that women's suffrage organizations were key in the promoting this justification for women's suffrage in the Western frontier states.

Braun and Kvasnicka (2013) offer a complementary political explanation for the leadership of the West: the relative scarcity of females in the West simply made it less costly for Western politicians to extend the franchise to women. Giving women the vote posed little risk to political stability since men greatly outnumbered women. Further, politicians may have viewed granting women suffrage as a means to attract more women to their states. Braun and Kvasnicka do find, consistent with their hypothesis, that the strongest predictor of the timing of full or presidential suffrage for women is the male-to-female ratio. They also find that states with higher percentages of females in the labor force enacted suffrage earlier, and states with higher fractions of nonwhites enfranchised women later. Unlike McCammon and Campbell (2001), Braun and Kvasnicka do not include controls for the strength and strategies of suffrage organizations across states, nor do they control for state-level prohibition laws. Braun and Kvasnicka discount the possible role for women's suffrage organizations by arguing that the West was "far less than [other regions] organized in terms of coordinated activities for securing the ballot" (408).

McCammon et al. (2001) expand on McCammon and Campbell (2001) to look at presidential as well as full suffrage and to consider a more extensive set of political and societal variables. McCammon et al. focus on how the strength and strategies of women's suffrage organizations varied across states, but they also point out that the success of these organizations' efforts depended on the willingness of political decision-makers to support the change in voting rights. This willingness was influenced by the potential political gains from extending suffrage to women, as well as attitudes about women in society. They find that states where women's suffrage organizations used separate spheres arguments, asserting that women would bring greater morality and reform to the political process, enacted suffrage earlier. States with a greater share of women in professions and attending college, both of which may indicate more progressive attitudes towards women, also enfranchised women sooner. Finally, states that had nullified liquor interest opposition by passing early prohibition laws were also more likely to enact suffrage earlier.

McConnaughy (2013) focuses on the role of political competition in the enactment of women's suffrage by the states. She finds that the extension of women's suffrage was more likely when there was a threat of third-party competition, as measured by the share of the vote received by third-party gubernatorial candidates and the share of third-party legislators in the state house (McConnaughy 2013, 223–24).

Like Braun and Kvasnicka (2013), McConnaughy (2013) finds that states with higher fractions of nonwhites were less likely to extend suffrage to women, perhaps because legislators in states with large percentages of nonwhites feared that enfranchising black women would erode white power. This race-based fear was evident during political debate about the Nineteenth Amendment. On June 4, 1919, the day the Senate voted to extend the right of suffrage to women, Senator Ellison D. Smith (D, SC) stated (58 Cong. Rec. 618, June 4, 1919):

> Those of us from the South, where the preponderance of the Negro vote jeopardized our civilization, have maintained that the fifteenth amendment was a crime against our civilization. Now, when a southern man votes for the Susan B. Anthony amendment he votes to enfranchise the other half of that race, and ratifies, not in a moment of heat and passion, what we have claimed was a crime, but in a moment of profound calmness and sectional amity he votes to ratify the fifteenth amendment and give the lie to every protestation that we have heretofore have made that the enfranchisement of the Negro men, unlimited, was a crime against white civilization.

While the empirical examinations of the timing of women's suffrage across the states do not produce a consensus explanation, none of them fully tests the competing hypotheses against each other. The general picture that emerges from these studies is that the extension of women's suffrage was the outcome of a political process. Male legislators and voters weighed the costs versus benefits of women's suffrage, and women's suffrage organizations took actions to increase the political benefits of supporting suffrage.

Conclusion

Economic theories of suffrage extension suggest that groups in power extend voting rights to promote their policy agendas or to capture votes from political rivals. Each of these factors played a role in the women's suffrage movement in the United States, although not necessarily at the same time and in the same way. Women's suffrage organizations sought to influence these political processes. Although some economists have discounted their contributions, these organizations were key to the success of the movement. They formed coalitions with other social movements and built a base of support that created electoral pressure to extend the vote to women. The empirical evidence also suggests that women were more likely to get the vote when men and other groups (such as liquor interests) had less to lose or when elections were close.

How did women having the vote change the political process? A number of scholars have taken advantage of the geographic variation in the timing of women's suffrage to look at the impact on public policy. Lott and Kenny (1999) and Miller (2008) both find a positive relationship between suffrage extension and public goods expenditures. Miller (2008) further shows that women gaining the vote led to reductions in child mortality, indicating that the increased public goods expenditures were being allocated in ways that improved child health. Moehling and Thomasson (2012) find that the timing of women's suffrage had a statistically significant impact on states' decisions to participate in the Sheppard-Towner program, which provided federal funds for public health education for mothers in the 1920s. States that were late-enactors of women's suffrage engaged more fully in this program perhaps because policymakers in these states sought to court the new women voters. Carruthers and Wanamaker (2015) find that suffrage led to an increase in public school expenditures.

Assessments of the longer term impact of the Nineteenth Amendment suggest that, despite generating a tremendous expansion in the electorate, the 1920 election did not lead to a seismic change in the political structure. Rather than revolution, the outcome of the election was a return to "normalcy" (Keyssar 2009, 175). Some politicians initially feared that women would vote as a bloc, but by the mid-1920s, it was clear that women's political interests were as varied as those of men. Moehling and Thomasson (2012) argue that the fate of the Sheppard-Towner program reflected this evolution of politicians' views on women's voting power. Politicians voted for the program in 1921 wanting to demonstrate their support for a "woman's issue." But by 1926, when the program came up for renewal, politicians no longer were worried about losing women's votes, and the program was repealed rather than renewed. The longer term legacy of the Nineteenth Amendment is discussed in the companion paper in this symposium by Cascio and Shenhav.

The Nineteenth Amendment also failed to provide universal suffrage for black women. The Southern states blocked black women from voting by using Jim Crow laws, just as they had long done to black men. Women's suffrage organizations refused to take up this cause, a stance consistent with the ways black women had been marginalized in the women's suffrage movement. As black women sought to join national women's organizations in the nineteenth century, they had been continually rebuffed. Suffrage supporters at a women's rights convention in Ohio in 1851 only very reluctantly let abolitionist Sojourner Truth take the floor to deliver her now-famous "Ain't I a Woman" speech, worrying that she would harm their cause (Buhle and Buhle 2005, 104–05). Even in the North, blacks were second-class citizens who were routinely excluded by whites. In one prominent counterexample, an 1895 speech in which she exhorted black women to take leadership, suffragist Josephine St. Pierre Ruffin (1895) called on white women to join with them, stating "[W]e are not drawing the color line; we are women, American women, as intensely interested in all that pertains to us as such as all other American women; we are not alienating or withdrawing, we are only coming to the front, willing to join any others in the same work and cordially inviting and welcoming any others to join us."

While whites mostly ignored Ruffin's (1895) appeal, new clubs for black women formed. Mary Church Terrell served as the first president of the National Association of Colored Women in 1896 (Flexner and Fitzpatrick 1996, 183). She was one of the few blacks allowed to speak at the segregated National American Women's Suffrage Association Convention in 1903, where she urged white women to "stand up not only for the oppressed sex, but also for the oppressed race" (Stanton et al. 1881, 5 and 106). In the first national suffrage parade held by the National American Woman Suffrage Association in Washington, D.C. in 1913, Terrell, along with Ida B. Wells, was forced to march with other blacks in the back so as not to upset Southern delegates, which Wells famously refused to do (Terborg-Penn 1998, 122–23).

After the passage of the Nineteenth Amendment, prominent women's suffrage organizations continued to refuse to fight for the rights of black women to vote when those rights were curtailed by Jim Crow laws like poll taxes and literacy tests. In 1921, Alice Paul refused to let Mary C. Talbert, the president of the National Association of Colored Women's Clubs, speak on black voting rights at the National Woman's Party convention. Paul defended her decision by asserting that this was about race, not women's rights (Cott 1984, 50–54). It would not be until 1965, with the passage of the Voting Rights Act, that *all* women could be said to have the right to vote.

References

Bertocchi, Graziella. 2011. "The Enfranchisement of Women and the Welfare State." *European Economic Review* 55 (4): 535–53.

Braun, Sebastian, and Michael Kvasnicka. 2013. "Men, Women, and the Ballot: Gender Imbalances and Suffrage Extensions in the United States." *Explorations in Economic History* 50 (3): 405–26.

Buhle, Mari Jo, and Paul Buhle, eds. 2005. *The Concise History of Woman Suffrage: Selections from the History of Woman Suffrage.* Champaign: University of Illinois Press.

Carruthers, Celeste K., and Marianne H. Wanamaker. 2015. "Municipal Housekeeping: The Impact of Women's Suffrage on the Public Education." *Journal of Human Resources* 50 (4): 837–72.

Cott, Nancy F. 1984. "Feminist Politics in the 1920s: The National Woman's Party." *Journal of American History* 71 (1): 43–68.

Doepke, Matthias, and Michèle Tertilt. 2009. "Women's Liberation: What's in It for Men?" *Quarterly Journal of Economics* 124 (4): 1541–91.

DuBois, Ellen Carol. 1987. "Outgrowing the Compact of the Fathers: Equal Rights, Woman Suffrage, and the United States Constitution, 1820-1878." *Journal of American History* 74 (3): 836–62.

Flexner, Eleanor, and Ellen Fitzpatrick. 1996. *Century of Struggle: The Woman's Rights Movement in the United States.* Cambridge: Belknap Press.

Geddes, Rick, and Dean Lueck. 2002. "The Gains from Self-Ownership and the Expansion of Women's Rights." *American Economic Review* 92 (4): 1079–92.

Grimes, Alan P. 1967. *The Puritan Ethic and Woman Suffrage.* New York: Oxford University Press.

Harper, Ida Husted. 2005. *The Life and Work of Susan B. Anthony (Volume 1 of 2): Including Public Addresses, Her Own Letters and Many from Her Contemporaries during Fifty Years.* Salt Lake City: Project Gutenberg. https://www.gutenberg.org/files/15220/15220-h/15220-h.htm.

Hooks, Janet M. 1947. *Women's Occupations through Seven Decades.* U.S. Department of Labor Women's Bureau Bulletin 218. Washington, D.C.: US Government Printing Office.

Jones, Ethel B. 1991. "The Economics of Woman Suffrage." *Journal of Legal Studies* 20 (2): 423–37.

Keyssar, Alexander. 2009. *The Right to Vote: The Contested History of Democracy in the United States.* New York: Basic Books.

Khan, B. Zorina. 1996. "Married Women's Property Laws and Female Commercial Activity: Evidence from United States Patent Records, 1790–1895." *Journal of Economic History* 56 (2): 356–88.

King, Brayden G., Marie Cornwall, and Eric C. Dahlin. 2005. "Winning Woman Suffrage One Step at a Time: Social Movements and the Logic of the Legislative Process." *Social Forces* 83 (3): 1211–34.

Klinghoffer, Judith Apter, and Lois Elkis. 1992. "'The Petticoat Electors': Women's Suffrage in New Jersey, 1776-1807." *Journal of the Early Republic* 12 (2): 159–93.

Lizzeri, Alessandro, and Nicola Persico. 2004. "Why Did the Elites Extend the Suffrage? Democracy and the Scope of Government, with an Application to Britain's 'Age of Reform.'" *Quarterly Journal of Economics* 119 (2): 707–65.

Llavador, Humberto, and Robert J. Oxoby. 2005. "Partisan Competition, Growth, and the Franchise." *Quarterly Journal of Economics* 120 (3): 1155–89.

Lott, John R., Jr., and Lawrence W. Kenny. 1999. "Did Women's Suffrage Change the Size and Scope of Government?" *Journal of Political Economy* 107 (6): 1163–98.

McCammon, Holly J. 2003. "'Out of the Parlors and into the Streets': The Changing Tactical Repertoire of the U.S. Women's Suffrage Movements." *Social Forces* 81 (3): 787–818.

McCammon, Holly J., and Karen E. Campbell. 2001. "Winning the Vote in the West: The Political Successes of the Women's Suffrage Movements, 1866-1919." *Gender and Society* 15 (1): 55–82.

McCammon, Holly J., and Karen E. Campbell. 2002. "Allies on the Road to Victory: Coalition Formation between the Suffragists and the Woman's Christian Temperance Union." *Mobilization: An International Quarterly* 7 (3): 231–51.

McCammon, Holly J., Karen E. Campbell, Ellen M. Granberg, and Christine Mowery. 2001. "How Movements Win: Gendered Opportunity Structures and U.S. Women's Suffrage Movements, 1866 to 1919." *American Sociological Review* 66 (1): 49–70.

McCammon, Holly J., Lyndi Hewitt, and Sandy Smith. 2004. "'No Weapon Save Argument': Strategic Frame Amplification in the U.S. Woman Suffrage Movements." *Sociological Quarterly* 45 (3): 529–56.

McConnaughy, Corrine M. 2013. *The Woman Suffrage Movement in America: A Reassessment.* New York: Cambridge University Press.

McDonagh, Eileen L., and H. Douglas Price. 1985. "Woman Suffrage in the Progressive Era: Patterns of Opposition and Support in Referenda Voting, 1910–1918." *American Political Science Review* 79 (2): 415–35.

Miller, Grant. 2008. "Women's Suffrage, Political Responsiveness, and Child Survival in American History." *Quarterly Journal of Economics* 123 (3): 1287–1327.

Moehling, Carolyn M., and Melissa A. Thomasson. 2012. "The Political Economy of Saving Mothers and Babies: The Politics of State Participation in the Sheppard-Towner Program." *Journal of Economic History* 72 (1): 75–103.

Nicholas, Kathryn A. 2018. "Reexamining Women's Nineteenth-Century Political Agency: School Suffrage and Office-Holding." *Journal of Policy History* 30 (3): 452–89.

Ruffin, Josephine St. Pierre. 1895. "Address to the First National Conference of Colored Women." Speech, First National Conference of Colored Women, Boston, July 29. https://www.blackpast.org/african-american-history/1895-josephine-st-pierre-ruffin-address-first-national-conference-colored-women/.

Stanton, Elizabeth Cady, Susan B. Anthony, Matilda Joslyn Gage, and Ida Husted Harper. 1881. *History of Woman Suffrage.* 6 Vols. Rochester: Susan B. Anthony.

Terborg-Penn, Rosalyn. 1998. *African American Women in the Struggle for the Vote, 1850-1920.* Bloomington: Indiana University Press.

Weiss, Elaine. 2018. *The Woman's Hour: The Great Fight to Win the Vote.* New York: Viking.

Journal of Economic Perspectives—Volume 34, Number 2—Spring 2020—Pages 24–48

A Century of the American Woman Voter: Sex Gaps in Political Participation, Preferences, and Partisanship since Women's Enfranchisement

Elizabeth U. Cascio and Na'ama Shenhav

The November 2020 presidential election will be the twenty-sixth in which American women will have been eligible to vote. The Nineteenth Amendment to the US Constitution, adopted a century ago this August, entitled women to cast ballots—for many in their first election—on November 2, 1920. Before 1920, 15 states had granted women equal voting rights, and an additional 24 states had granted partial voting rights (Keyssar 2009; Lott and Kenny 1999). The scope of partial voting rights varied widely across states and municipalities, covering local school board elections to the national presidential election. But a constitutional guarantee that sex could not be used as a basis of exclusion from the vote represented the crowning achievement—the "sacred right to the elective franchise," as laid out in the Declaration of Sentiments at the celebrated Seneca Falls convention of 1848.

But paradoxically, there was a great deal of speculation back in 1920 that women might not have much political influence even with the right to vote. As the story went, women would be inclined to "duplicate" the male vote, if they turned out at the polls at all. Some argued that any other outcome would be too disruptive to household harmony and also disrespectful of the "separate spheres" that men and women had historically occupied. As late as 1940, pollster George Gallup mused, "How will [women] vote on election day? Just exactly as they were told the

■ *Elizabeth U. Cascio is Associate Professor of Economics at Dartmouth College, Hanover, New Hampshire, and Research Associate at the National Bureau of Economic Research, Cambridge, Massachusetts. Na'ama Shenhav is Assistant Professor of Economics at Dartmouth College, Hanover, New Hampshire. Their email addresses are elizabeth.u.cascio@dartmouth.edu and naama.shenhav@dartmouth.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.2.24.

night before" (as quoted in Berinsky 2006, 506). Others believed that other aspects of a woman's identity—her social class, race, or immigrant status—would be more critical than her sex to her political choices. Because women were similarly distributed across these other groups, so too would be women's votes—and public policy would be little affected.

A growing literature looking at a number of outcomes challenges this narrative. A series of studies using area-by-time variation has examined the impacts of women's enfranchisement on levels of state and municipal spending (Lott and Kenny 1999; Miller 2008), the distribution of spending across priorities, like public health and education (Miller 2008; Moehling and Thomasson 2012; Carruthers and Wanamaker 2015; Kose, Kuka, and Shenhav 2019), electoral outcomes (Morgan-Collins forthcoming), and downstream impacts on human capital in the short and long term (Miller 2008; Kose, Kuka, and Shenhav 2019). The findings are consistent with survey, lab, and field evidence from a variety of settings and time periods suggesting women place higher priorities on child welfare and redistribution (for reviews, see Duflo 2012; Croson and Gneezy 2009). Thus, the evidence suggests that the women's vote translated into real impacts on policy and social welfare, even at a time—as we will show—that women participated less in the electoral process relative to the present day.

Our goal in this paper is not to revisit the immediate impacts of the Nineteenth Amendment. Rather, we aim to describe how women as political actors have evolved over the past century, a period when women have had de jure—even if not always de facto—full voting rights. We will focus on three sets of outcomes: *political participation* is involvement in the political process, *issue preferences* are preferences over policy outcomes, and *partisanship* is identification with specific political parties and candidates. We posit that women's potential for political influence is an increasing function of their participation and how much both their issue preferences and partisanship differ from men's. When relatively high political participation intersects with relatively different issue preferences and there is sufficient party polarization along divisive preferences, the so-called "women's vote" may become pivotal to the candidates elected and the policies enacted.

To describe the evolution of the female voter in the United States, we bring together data from a variety of data sources. For the elections immediately after 1920, we do not have survey data on voting patterns by sex, but we can draw inferences about voting behavior of women from overall changes in voter turnout. We also compile a range of survey data from Gallup polls and other sources on patterns of men and women voting back to 1940. Wolbrecht and Corder (forthcoming) contemporaneously use a similar scope of data to analyze time trends in the sex gap across a wide breadth of voting-related outcomes. To the extent that our analyses overlap, we find consistent results. However, the scale and detail of our data allow us to explore the drivers of change in data-intensive, novel ways. Paralleling analyses of the growth in women's labor force participation over time (for example, Goldin 1990; Bailey 2006), we bring a new focus on the contributions of cohort- and time-specific factors in shaping voting outcomes.

We arrive at two key sets of findings. First, we show that the female-male gap ("sex gap") in voter turnout grew substantially over the last 80 years, from a deficit of almost 10 percentage points in 1940 to a surplus of over 4 percentage points in the 2016 election. This shift has been driven primarily by an increase in women's relative turnout across cohorts, which we find is associated with the accompanying rise in education, particularly high school graduation. Second, we show that the sex gap in identification with the Democratic Party rose from roughly parity in the late 1940s to almost 12 percentage points in 2017. This shift in partisanship permeated all cohorts, but did not coincide with a significant shift in issue preferences. We present survey evidence consistent with observations by political scientists (for example, Layman and Carsey 2002; Gillion, Ladd, and Meredith 2018) that party polarization across recent decades has contributed to a widening gulf in party affiliation by sex.

While we will not attempt to provide new estimates of the real policy and economic impacts of the female voter, we provide new descriptive evidence on when and how women's potential for political influence changed since the Nineteenth Amendment. However, we urge caution in drawing strong inferences from these descriptive patterns. After all, despite the higher voter turnout rates and greater Democratic partisanship of women as a whole, Republican Donald J. Trump won the 2016 presidential election. Post-election analyses have emphasized the role of divisions within women, particularly by race, which serves as a reminder that the notion of a "woman voter" is a vast simplification, as is the notion of a "man voter."[1] Exploring long-term trends in women's relative political behavior in subpopulations defined by race, education, marital status, geography, and so on is beyond the scope of the present paper, but an important area for future research.

Political Participation

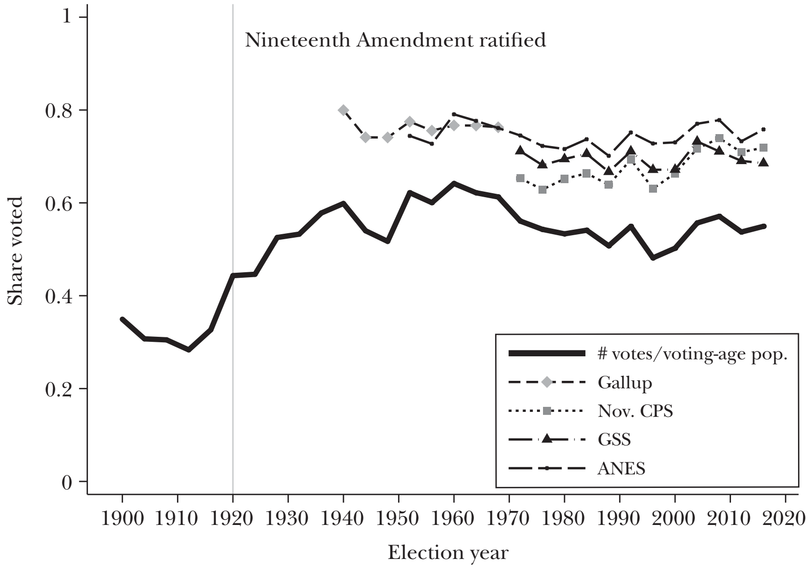
We begin with a new investigation of women's relative political participation since 1920. We focus on voter turnout in presidential elections, the measure of turnout that can be most consistently observed over the longest time horizon. Presidential elections also have the highest voter turnout, allowing us to observe the frontier of voter turnout for men and women alike. We consider the extent to which other participation metrics for women moved along with their turnout in a supplemental analysis to follow.

National Trends

There is not any direct data on voter participation of men and women in the 1920 election, or for the several elections that follow. Thus, researchers have sought to infer the voter participation rates of women based on overall voter turnout.

[1] See Cassese and Barnes (2018) for analysis of the 2016 election and broader discussions in Wolbrecht and Corder (forthcoming).

Figure 1
Voter Turnout in US Presidential Elections: Survey and Administrative Data, 1900–2016



Source: The numerator of the series represented by the solid black line is the US presidential vote count, constructed by the authors from state-level vote tallies available at http://uselectionsatlas.org. The denominator of this series is the US voting-age population (ages 18 and up for 1972 forward+ and ages 21 and up in all earlier years) for the subsample of states in a region with election returns, estimated from Decennial Census (1900–2000) and American Community Survey (ACS) (2005–2016) Public Use Microdata Samples (Ruggles et al. 2019). We compile Gallup microdata from polls conducted from 1940–1970 by the Gallup Organization and November CPS Voter Supplement microdata (for 1972; from US Census Bureau 1992) and IPUMS CPS (for 1976–2016; from Flood et al. 2018). General Social Survey data are drawn from the General Social Survey 1972–2018 Cross-Sectional Cumulative Data file (Release 1) (for 1972–2016; from Smith et al. 2019) and American National Election Studies data from the American National Election Studies Time Series Cumulative Data File (for 1952–2016; from American National Election Studies 2019). See online data Appendix.
Note: We weight statistics from the Gallup microdata using weights that we construct from the census, which adjust Gallup demographics to the year × region × education × race × sex × birth cohort level. (Birth cohorts are defined as described later in the paper.) We weight statistics from the November CPS, General Social Survey, and American National Election Studies using survey-provided weights. All weights are re-normed so as to average to one within each survey-year.

The bold line in Figure 1 plots national voter turnout in presidential elections based on aggregation of state-level vote counts. We divide the number of votes cast in a presidential election nationally by an estimate of the total voting-age population in states with election returns.[2] We thus allow the denominator to include both

[2] The Twenty-Sixth Amendment extended the franchise to 18–20 year-olds in 1971. Thus, the voting-age population consists of persons aged 21 and older in elections through 1968 and persons aged 18 and older in 1972 and later.

men and women, even during the pre-1920 period when women were generally not eligible to vote in presidential elections. We also include noncitizens in the denominator, since a citizenship question was not consistently asked in the census from 1900 forward.

Some states granted women the right to vote in presidential elections prior to 1920: six states in 1912 and twelve states in 1916.[3] However, these states were concentrated in the sparsely populated West and therefore comprised a small share of the total population. If women voted at the same rate as men, women's suffrage should then have led to a near-doubling of voter turnout. Measured against this standard, women entered the electorate slowly. Between 1916 and 1920, voter turnout increased by only 35 percent, from 32.7 to 44.3 percent. However, over the next 20 years, voter turnout continued to increase basically unabated, reaching 59.8 percent in 1940.

The descriptive pattern from overall voting totals is consistent with other indirect methods. Using variation in the timing of state suffrage initiatives in addition to ratification of the Nineteenth Amendment, Kose, Kuka, and Shenhav (2019) find that women's suffrage increased voter turnout in the short term by 56 percent. Taking a Bayesian approach to data from ten states, Corder and Wolbrecht (2016) also find substantial, and generally shrinking, sex gaps in turnout across the five presidential elections from 1920 to 1936.

To describe the evolution of the female US voter in more detail, we turn to survey data. We provide an overview of the data here, with further details in a Data Appendix available with this paper at the journal website. Scholars of American politics interested in long-term trends in political behavior typically rely on data from the American National Election Studies or the General Social Survey. These surveys are detailed—and we will also use them—but have sample sizes of only around several thousand per election and start later than ideal for our purposes (1952 and 1972, respectively). To extend backward in time and obtain more data for the 1950s and 1960s, we turn to historical polling data collected by the Gallup Organization's American Institute of Public Opinion (AIPO). These data have been used on a limited but growing basis by economists (for example, Fogli and Veldkamp 2011; Kuziemko and Washington 2018; Farber et al. 2018). The standard question on voter participation first prompts respondents about whether they are certain they voted (few say no), then asks about candidate chosen. The Gallup data begin in 1936, but we start our series in 1940, the first year in which respondents are asked about their education.

[3] These included Washington, California, Idaho, Utah, Wyoming, and Colorado, each of which passed full suffrage by 1912; and Oregon, Arizona, Montana, Nevada, Kansas, and Illinois, which either passed full or presidential suffrage by 1916. See Kleppner (1982) and Corder and Wolbrecht (2016) for discussions of turnout in these early elections. An additional 16 states, including North Dakota, South Dakota, Nebraska, Kansas, Oklahoma, Texas, Minnesota, Iowa, Wisconsin, Michigan, Indiana, Ohio, Tennessee, New York, Rhode Island, and Maine, passed presidential or full suffrage prior to the adoption of the Nineteenth Amendment. See Teele (2018) for a mapping of the timing of these and other voting rights (like voting in primary elections) that were passed during this period.

With the addition of data from the November Voter Supplement of the Current Population Survey (CPS), starting in 1972, we have microdata on voting in presidential elections that span nearly 80 years: 1940 to 2016.[4] Focusing on Gallup polls conducted within two years after a given presidential election, we obtain sample sizes at least an order of magnitude higher than those available in the American National Election Studies and the General Social Survey. The November CPS similarly offers large samples that allow us to explore the drivers of trends in the sex gap in data-intensive ways.

Figure 1 includes national trends in voter turnout based on these four survey sources: Gallup polls, the American National Election Studies, the General Social Survey, and the November CPS. We use weights provided in the last three surveys to generate nationally representative statistics. For the Gallup data, we create weights from census microdata to adjust Gallup demographics to match the distribution of the population across cells defined by year, region, education, sex, and birth cohort. Weighting of the Gallup data is especially important because the sampling approach used by Gallup into the 1950s had a goal of representing the "engaged public," rather than the population at large (Berinsky 2006). Thus, the unweighted Gallup data in early years will underrepresent those with less education, the South, the nonwhite population, and women.[5]

Regardless of the survey source or year, self-reported voter turnout is consistently higher than the administrative measure—a well-known feature of self-reports of voting (for example, Bernstein, Chadha, and Montjoy 2001; Ansolabehere and Hersh 2012). The administrative and survey series nevertheless move together, suggesting that the survey data capture important margins of change in voter turnout from election to election. There are also less pronounced but still noticeable differences in levels of voter turnout across surveys. American National Election Studies and Gallup data consistently produce higher turnout estimates than the General Social Survey and November CPS. However, focusing on the *sex gap* in turnout—our measure of interest—will eliminate survey effects that are the same across sex. In addition, a recent validation study (Ansolabehere and Hersh 2012) suggests that women have similar levels of misreporting as men in the American National Election Studies.

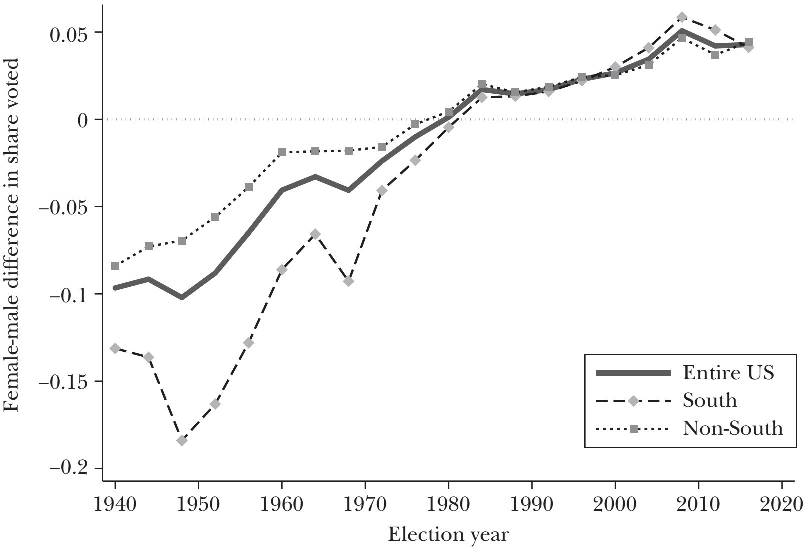
The Sex Gap in Participation

The solid line in Figure 2 plots the difference between female and male turnout combining data from our four survey sources. Overall, the figure shows a stunning story of change. While aggregate voter turnout varied from election to election, the sex gap in turnout was roughly constant at about 10 percentage points during the 1940s and 1950s. In other words, women were consistently about 10 percentage

[4] As we mentioned above, contemporaneous work by Wolbrecht and Corder (forthcoming) assembles a similar dataset to ours; however, while we pool together information from overlapping data sources to gain additional precision, Wolbrecht and Corder analyze each data source separately.

[5] Application of the weights tends to bring these characteristics in line with national averages, as shown in Appendix Figure 1. See online Data Appendix for a complete description of the weights and their creation.

Figure 2
Sex Gap in Voter Turnout, US Overall and by Region: Pooled Survey Data, 1940–2016



Source: Survey data pool the Gallup, November CPS, General Social Survey, and American National Election Studies series described in the notes to Figure 1.
Note: Statistics are weighted by survey-provided weights (for the CPS, General Social Survey, and American National Election Studies) or author-constructed weights (for Gallup), with all weights re-normed to average to one within each survey-year. More details are in the online data Appendix. The figure plots the difference in estimated voter turnout rates between women and men by year, nationally, and separately by region, with South representing the southern census region.

points less likely to vote than men. The gap dramatically narrowed thereafter, however, reaching about a 3 percentage point deficit for women in 1964. Though the gap re-expanded somewhat in 1968, women's voter turnout rates fell relatively less over the 1970s than men's, enough that women's and men's turnout basically reached parity by 1980. Women's voter turnout continued to gain in relative terms after 1980. In the last three presidential elections, women have been about 4 to 5 percentage points *more* likely to vote than men. Because women make up more than half of the voting-age population, they became the majority of voters earlier—in 1960, according to our data.

American women thus appear to have become increasingly comfortable exercising their right to vote. Is this pattern of convergence and eventual female dominance in political participation apparent in other metrics? Table 1 summarizes a series of political interest and mobilization variables available both in the 1950s and more recently in the American National Election Studies. The sex gap in some—but not all—of these measures shows a similar pattern as for voter turnout. For example, women on average used to care less about which party won an election and were less

Table 1

Trends in the Sex Gap in Voter Participation: Measures from the American National Election Studies

	1950s		*2010s*	
	Mean	*Sex gap*	*Mean*	*Sex gap*
A. Turnout				
Voted in last presidential election	0.735	−0.112	0.746	0.017
B. Political interest				
Cares a lot about which party wins presidential election	0.649	−0.047	0.813	0.000
Somewhat or very interested in elections	0.702	−0.064	0.844	−0.019
Very interested in elections	0.334	−0.073	0.443	−0.066
C. Mobilization				
Tried to influence someone's vote	0.278	−0.121	0.435	−0.041
Displayed candidate button/sticker during campaign	0.155	−0.060	0.139	−0.008
Donated money to party candidate during campaign	0.072	−0.031	0.120	−0.024
Attended political meetings/rallies during campaign	0.069	−0.020	0.063	−0.008
Worked for party or candidate during campaign	0.032	−0.011	0.033	−0.010

Source: Data are from the American National Election Studies (American National Election Studies).
Note: Data for the 1950s pertain to the 1952 and 1956 elections (with the exception of the variable "displayed candidate button/sticker during campaign," which is only available for 1956). Data for the 2010s pertain to the 2012 and 2016 elections. Statistics are weighted by American National Election Studies sampling weights. Sex gap is the female-male difference in the outcome.

interested in elections; they also used to be less likely to try to influence someone's vote or to display campaign paraphernalia. Sex differences in these attitudes and behaviors are now largely gone. However, sex gaps in rarer measures—being "very interested" in elections, making political donations, attending campaign rallies, and working for campaigns—have remained largely unchanged over time.

These findings thus seem to suggest a relatively dramatic narrowing of the sex gap in mass, but not extreme, political participation. At the same time, however, women's participation as elected officials—another extreme participation metric—*has* increased over time, though the sex gap still strongly favors men.[6] For example, nearly one-quarter of current members of the Senate and House of Representatives are women, compared to 10 percent following the 1992 election ("the year of the woman") and less than 3 percent in the early 1950s. These statistics closely track the increasing propensity of women to run in a congressional primary (Lawless and Pearson 2008).

[6] The persistence in this gap could reflect gender differences in preferences or in socialization around political careers. A recent survey of college students suggests that women are less likely to have political ambitions but are also less likely to have received parental encouragement to run for political office (Lawless and Fox 2013). See also Wasserman (2018), which shows that women are less likely to run again for political office after a loss.

Sex, Race, and Persistence of Limits on the Franchise after 1920

Not all women actually gained the franchise in 1920: in particular, black women in the South were largely excluded. Although black men were granted the right to vote after the Civil War via ratification of the Fifteenth Amendment, southern states subsequently designed a series of electoral devices—poll taxes and literacy tests at voter registration in particular—to disenfranchise them (Keyssar 2009; Valelly 2004). Historical evidence suggests that these devices similarly limited southern black women's entrée into voting booths in 1920; for example, voter turnout estimated as a ratio of votes cast to voting-age population shows a weak response to the Nineteenth Amendment in the South, relative to other regions like the Northeast and the Midwest. However, poll taxes were eliminated by a combination of state action and ratification of the Twenty-Fourth Amendment in January 1964, and literacy tests were removed via passage of the Voting Rights Act of 1965. Civil rights activism may have also helped register southern blacks and get them to the polls, even before structural barriers to participation were removed.[7]

How might this history have contributed to the evolution of the national sex gap in voter turnout? The answer to this question will depend on whether there were sex differences in the efficacy of both the disenfranchising measures and the remedies. Anti-suffragists in the South worried that it would be more difficult to use the tactics that had been applied to black men to staunch the vote of black women. One Mississippi senator said: "We are not afraid to maul a black man over the head if he dares to vote, but we can't treat women, even black women, that way" (as quoted in Keyssar 2009, 169). By this reasoning, southern black women would have been more likely to vote than southern black men, potentially narrowing the sex gap in turnout in the South relative to the rest of the country even early in the period. Contrary to this hypothesis, Figure 2 shows that the sex gap was actually much larger in 1940—and male-female convergence in voter turnout thereafter much more dramatic—in the southern census region. Further exploration of the data shows that the sex gap in voter turnout in the South from the 1940s through the early 1960s was roughly the same for whites and nonwhites, suggesting that forces that were unique to the region—but not necessarily to any particular race—contributed to the marked closure in the sex gap over this same period.

[7] See online Appendix Figure 2 for the time series of turnout by region. Voter turnout in the South did not converge to that in the rest of the country until after poll taxes and literacy tests were removed. See Cascio and Washington (2014) and Filer, Kenny, and Morton (1991) for causal evidence on this link using historical voting records and geographic variation in the black share of the population. We see similar patterns in our data, which affords us voting information by race (see online Appendix Figure 3).

Preferences over Policies and Parties

Background: Theory and Evidence

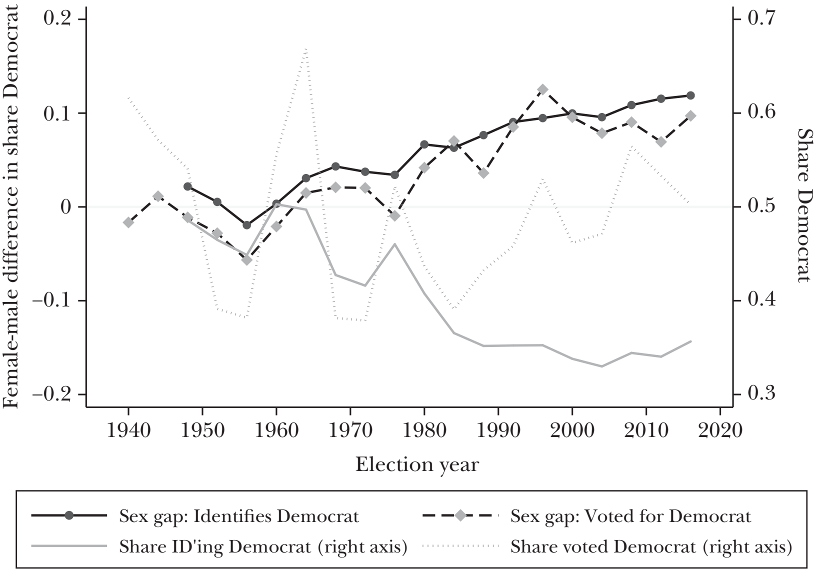
Over the past 80 years, women's political mobilization has not only steadily converged with men's, but has overtaken it. Does this mean that women have been increasingly influential for political outcomes over time? Theoretical models suggest that the implications of women's political participation for policy may depend on not just women's turnout and policy preferences relative to men's (Downs 1957; Cox and McCubbins 1986), but also the nature of political competition. If politicians are able to implement their preferred policies (Osborne and Slivinski 1996; Besley and Coate 1997), for example, having an impact on policy would require voting for candidates with shared ideology, which could be captured by party affiliation.

In the years leading up to and following passage of the Nineteenth Amendment, women had different issue preferences than men—valuing policies that benefited children and public health and welfare and more government spending—but not dramatically different party alignment. Studies using variation in the timing of state laws enabling women to vote in state and local elections prior to the Nineteenth Amendment find impacts on policy and real economic outcomes at the federal, state, and local levels that move toward women's preferences (Lott and Kenny 1999; Miller 2008; Moehling and Thomasson 2012; Carruthers and Wanamaker 2015; Kose, Kuka, and Shenhav 2019). However, Corder and Wolbrecht (2016) show that, in the first five presidential elections in which women could vote, their votes went toward parties much the same way as those of the men in their state. This seeming contradiction may be explained by the political environment: historically, the two major parties were not well sorted on the dimensions of public opinion along which men and women tend to differ systematically (Gillion, Ladd, and Meredith 2018). While this was particularly the case before the 1930s (Gerring 1998), our data suggest this pattern held as recently as the 1970s, as argued elsewhere by political scientists.

What has happened over the past 50 to 60 years? Literature from across the social sciences suggests that the sex gap in public opinion on various issues has been fairly stable in the face of dramatic social and cultural change (Shapiro and Mahajan 1986; DiMaggio, Evans, and Bryson 1996; Bolzendahl and Myers 2004; Clark 2017). And yet, the sex gap in party identification has not been stable: instead, there has been a dramatic relative shift of women toward the Democratic Party as men have increasingly been drawn toward the Republican Party (Kaufmann and Petrocik 1999; Edlund and Pande 2002; Box-Steffensmeier, De Boef, and Lin 2004; Gillion, Ladd, and Meredith 2018). Other advanced industrialized countries also experienced a relative shift of women toward more liberal political parties over the latter half of the twentieth century (Inglehart and Norris 2000).

In this section, we consider the evolution of the sex gap in party affiliation and in policy preferences, and then ask: how can the sex gap in party affiliation have widened without a change in the sex gap in preferences?

Figure 3
The Sex Gap in Democratic Partisanship: Pooled Survey Data, 1940–2016



Source: Microdata on Democratic Party identification are from the replication archive of Gillion, Ladd, and Meredith (2018) (for 1953–2012), Gallup polls spanning 1948–1952 and 2013–2017 that we collected from the Roper Center, the American National Election Studies Time Series Cumulative Data File (1948–2016), and the General Social Survey Cross-Sectional Cumulative Data 1972–2018 (Release 1). Microdata on voting for the Democratic presidential candidate are from Gallup polls spanning 1940–1970, the American National Election Studies Time Series Cumulative Data File (1948–2016), and the General Social Survey Cross-Sectional Cumulative Data 1972–2018 (Release 1).
Note: Statistics are weighted by survey-provided weights (for the General Social Survey and American National Election Studies) or author-constructed weights (for Gallup), with all weights re-normed to average out to one within each survey-year.

The Sex Gap in Party Affiliation

Our core analysis of the sex gap in party affiliation is based on polling microdata, primarily from the Gallup Organization, on party identification spanning from 1953 to 2012, generously provided by Gillion, Ladd, and Meredith (2018). For consistency with our study of the sex gap in voter turnout and in party of the candidate chosen, we limit attention to polls taken within two years after an election, summarizing these polls with an election-year average. Applying this constraint, we use Gallup data to extend the series both backward in time to the 1948 election and forward in time to the 2016 election (as described in the online Data Appendix).

Figure 3 shows national trends in the sex gaps in identification with the Democratic Party and, for comparison, in vote share for the Democratic candidate. Estimated population shares voting for the Democratic candidate and identifying with the Democratic Party are shown for context (right axis).[8] The voting series is naturally punctuated by election years with Democratic victories, but overall, there is a clear reduction in Democratic Party identification between the 1960s and early 1980s, driven by political realignment in the South (Kuziemko and Washington 2018).

The figure shows that women have been increasingly more likely to affiliate with the Democratic Party—or rather, less likely to leave the Democratic Party (Kaufmann and Petrocik 1999)—than men. While a divergence in party preferences of American women and men emerged in the 1960s, it took off starting in the 1980s.[9] Following the 2016 election, women were almost 12 percentage points more likely than men to consider themselves Democrats, compared to a sex gap hovering around zero in the late 1940s and 1950s. Past work using the American National Election Studies (Kaufmann and Petrocik 1999), other polling data (Box-Steffensmeier, De Boef, and Lin 2004), or the same polling data used here but with different weighting (Gillion, Ladd, and Meredith 2018) has also documented an increasing partisan sex gap in the United States, though over shorter time horizons. While we focus on average gaps between men and women, recent work has shown that these gaps are often less pronounced among white voters than nonwhite voters (for example, Cassese and Barnes 2018).

The Sex Gap in Issue Preferences

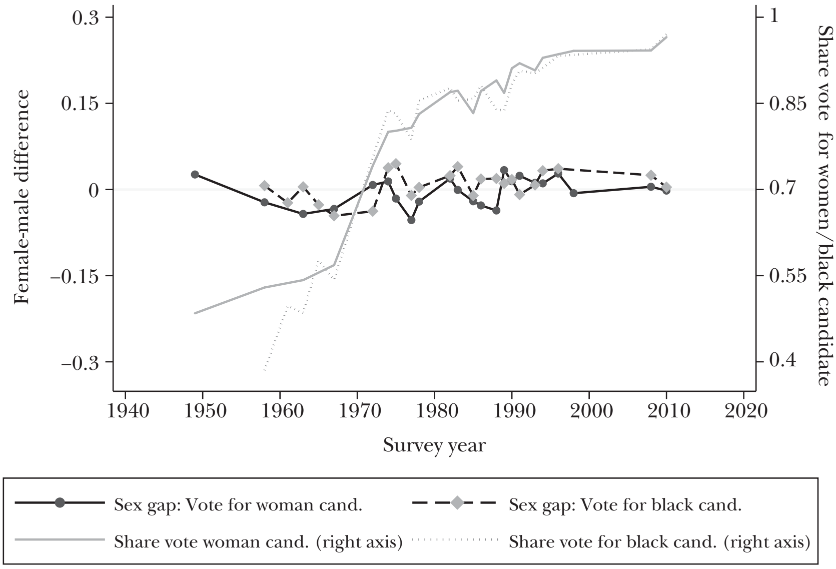
Political parties and political preferences do not necessarily align. There are no large-scale microdata asking consistent public opinion questions over the same time frame as represented in Figure 3. However, every few years starting in the 1940s, Gallup polls fielded questions concerning traits of hypothetical presidential candidates: for example, whether one would vote for a qualified woman if she were the nominee of one's party (starting in 1949), or for a qualified black man (starting in 1958). To these data, we add responses to a similar set of questions from the General Social Survey in more recent years (until 2010). The answers provide some insight into both the magnitude of social change over the period of interest and sex differences in reactions to it.

Even if the answers only represented changes in social desirability bias, the scope of social change represented in Figure 4 is breathtaking: the share of the population stating they are willing to vote for a female president rose from 47 percent in 1949 to 96 percent in 2010. Growth in the share of the population willing to vote

[8] National statistics on Democratic vote share from Gallup, General Social Survey and American National Election Studies data map fairly well to statistics based on historical voting records, though survey reports tend to favor the winning candidate, as shown in online Appendix Figure 4. Again, to the extent that this tendency is the same across sex, survey-based measures of the sex gap in partisanship should be representative.

[9] Online Appendix Figure 5 shows the parallel series for the Republican Party. Men are now more likely to identify as Republicans than women. However, unlike in the Democratic case, the population share identifying as Republicans is not that different today than in 1948. Unlike in the case of voter turnout, moreover, there are no significant regional differences in trends in the sex gap in Democratic Party identification, as shown in online Appendix Figure 6.

Figure 4
Sex Gap in Political Opinion: Preferences over Presidents, Pooled Survey Data, 1948–2010



Source: Microdata are from Gallup polls (1948–1969) and the General Social Survey Cross-Sectional Cumulative Data 1972–2018 (Release 1).
Note: Statistics are weighted by survey-provided weights (for the General Social Survey) or author-constructed weights (for Gallup), with weights re-normed to average to one within each survey-year.

for a black president has been even more striking, rising from 38 percent in 1958 to 97 percent in 2010. Yet the sex gaps in both measures have bounced around zero, showing no clear trend; indeed, men have more often than not exhibited *greater* support for the idea of a female president. This is consistent with existing evidence that suggests that women have historically not used the franchise to advance their own political or economic interests as a sex.

Table 2 presents a mixed pattern of changes in the sex gap in views on various policy topics reported in the General Social Survey from 1977 to 1986 and 2007 to 2016, the earliest and latest ten-year spans with consistent responses to our questions of interest. We summarize responses to 25 preference elicitations with seven indices, which are calculated as the mean of responses in a particular area (coded such that higher values always indicate more progressive views).[10]

[10] These questions and groupings strongly overlap with the questions and categories in DiMaggio, Evans, and Bryson (1996). Not all questions are asked in all years, but all of the questions in the indices appear both in the early and later periods (see online Appendix Table 1). Additional survey evidence from the American National Election Studies broadly confirms the patterns discussed here, in some cases with polls reaching back to the 1950s and 1960s (for results and details, see online Appendix Table 2).

Table 2

Trends in the Sex Gap in Issue Preferences: Evidence from the General Social Survey

	1977–1986		*2007–2016*	
	Mean	*Sex gap*	*Mean*	*Sex gap*
Voted in last presidential election	0.701	–0.009	0.698	0.024
Identifies as a Democrat	0.388	0.050	0.325	0.072
Voted for Democrat in last presidential election	0.418	0.047	0.554	0.112
Sexuality Attitudes Index	***0.281***	***–0.066***	***0.512***	***0.022***
Homosexual relations not wrong	0.137	–0.004	0.457	0.097
Okay to have sex before marriage	0.392	–0.112	0.569	–0.048
Criminal Justice Index	***0.277***	***0.069***	***0.376***	***0.049***
Courts too harsh	0.032	–0.007	0.162	–0.039
Should need gun permit	0.723	0.125	0.737	0.110
Oppose death penalty for murder	0.248	0.100	0.351	0.088
Abortion Attitude Index	***0.646***	***–0.024***	***0.613***	***–0.033***
Abortion if serious defect	0.816	–0.013	0.739	–0.032
Abortion if married + don't want more	0.428	–0.040	0.458	–0.048
Abortion if mom health at risk	0.904	–0.022	0.879	–0.010
Abortion if very poor	0.475	–0.019	0.440	–0.023
Abortion if pregnant from rape	0.829	–0.022	0.772	–0.044
Abortion if single + don't want to marry	0.437	–0.027	0.417	–0.025
Women's Public Roles Index	***0.710***	***0.007***	***0.791***	***0.032***
Disagree women not suited to politics	0.565	0.018	0.751	0.041
Vote woman president	0.842	–0.020	0.951	0.003
Family Gender Roles Index	***0.481***	***0.097***	***0.697***	***0.099***
Disagree woman should stay home	0.453	0.043	0.676	0.050
Agree mom working doesn't hurt kids	0.574	0.133	0.743	0.122
Disagree pre-K kids suffer if mom works	0.419	0.121	0.676	0.128
Progressive Government Index	***0.307***	***0.026***	***0.323***	***0.029***
Govt. should help poor	0.311	0.022	0.301	0.035
Govt. should help sick	0.466	–0.006	0.478	0.043
Govt. should help blacks	0.181	–0.001	0.189	0.009
Govt. should equalize wealth	0.306	0.035	0.321	0.030
Race Equality Index	***0.533***	***0.022***	***0.569***	***0.026***
Race gap not due to ability	0.780	0.041	0.905	–0.007
Race gap due to access	0.524	0.003	0.463	0.023
Race gap not due to motivation	0.390	0.023	0.526	0.026
Race gap due to discrimination	0.437	0.027	0.372	0.056

Source: Data are from the General Social Survey (General Social Survey).
Note: Statistics are weighted by General Social Survey sampling weights. The years in the column headers refer to survey years for preferences and election years for voting outcomes. For voting outcomes shown in the first two rows of the table, we also include the 2018 General Social Survey in which individuals report on voting behavior in the 2016 election. Sex gap is the female-male difference in the outcome.

In the earliest decade, 70 percent of respondents reported voting in the last election, with an immaterial gap across sexes; and 39 and 42 percent reported identifying with or voting for a Democrat, with a 5 percentage point sex gap favoring women. The sex gap in issue preferences varied in size and direction. The largest absolute sex gaps were in sexuality attitudes, where women espoused more conservative views (owing to less approval of premarital sex), and in the criminal justice and family gender roles index, where women were more progressive. There were somewhat smaller gaps in abortion attitudes and in the progressive government and race equality indices and opinions on the women's public roles index.

Over the next four decades of the survey, women's voting rates and propensity to identify or vote for a Democrat increased relative to men's by 3, 2, and 7 percentage points, respectively. In terms of preferences, we find a striking increase in the gap in the sexual attitudes index (9 percentage points), which includes more favorable views towards the gay community (Fernández, Parsa, and Viarengo 2019) as well as towards premarital sex. But otherwise, the changes are minor, with inconsistent signs. To a limited degree, women have become relatively more supportive of women's public roles and less supportive of reform of the criminal justice system. We do not see any meaningful change in relative views towards abortion, racial equality, support for government services (although women's relative support for government services to the sick does rise), or perceptions of the role of women as mothers first. These results align with previous research over somewhat shorter time horizons showing little movement in the sex gap in policy preferences (Clark 2017; DiMaggio, Evans, and Bryson 1996).

Party Polarization as Reconciliation

How can the sex gap in party affiliations have widened without an underlying change in preferences? Gillion, Ladd, and Meredith (2018) posit that the rise in Democratic identification among women since the 1970s represents a change in *sorting* across political parties, influenced by the increasing party polarization, driven by elites, and by growing public awareness of that polarization (Carsey and Layman 2006). In support of this hypothesis, they show in data from the American National Election Studies that the gender partisan gap is larger among those who are college-educated and more aware of the polarization across parties. They also show that the weight placed on social welfare and other preferences in the partisan identification decision has increased. Thus, they argue that (1) more educated groups would be in the best position to sort into political parties based on each party's current positions, and (2) changes in these issue weights would be predicted to induce a larger response by women, given existing gaps in preferences.

In a similar spirit, we turn to the General Social Survey to investigate the scope of changes in polarization over this period and their relevance for the sex gap in partisanship. In particular, we examine the total change in the gap in policy preferences between individuals that identify as Democrats and Republicans ("party gap"), within and across sexes. This reduced-form approach is purely descriptive, but provides a transparent look into these patterns over time in a wide variety of

Table 3
Trends in Party Polarization within and across Sex: Measures from the General Social Survey

	1977–1986		*2007–2016*	
	Dems.	*Party gap*	*Dems.*	*Party gap*
A. Women				
Voted in last presidential election	0.723	−0.079	0.824	−0.022
Voted for Democrat in last presidential election	0.739	0.660	0.944	0.839
Sexuality attitudes index	0.257	0.068	0.581	0.217
Criminal justice index	0.330	0.075	0.479	0.210
Abortion attitude index	0.628	−0.021	0.681	0.213
Women's public roles index	0.721	0.073	0.847	0.104
Family gender roles index	0.524	0.030	0.786	0.096
Progressive government index	0.403	0.214	0.464	0.319
Race equality index	0.551	0.051	0.644	0.144
B. Men				
Voted in last presidential election	0.734	−0.078	0.770	−0.060
Voted for Democrat in last presidential election	0.736	0.681	0.912	0.837
Sexuality attitudes index	0.313	0.057	0.539	0.147
Criminal justice index	0.263	0.065	0.440	0.225
Abortion attitude index	0.641	−0.029	0.679	0.139
Women's public roles index	0.687	0.018	0.813	0.097
Family gender roles index	0.399	−0.006	0.680	0.081
Progressive government index	0.381	0.189	0.450	0.321
Race equality index	0.505	0.008	0.617	0.145

Source: Data are from the General Social Survey (General Social Survey).
Note: Statistics are weighted by General Social Survey sampling weights. The years in the column headers refer to survey years for preferences and election years for voting outcomes. "Dems." refers to individuals that report identifying as a Democrat, strongly or not strongly. "Party gap" is the difference between views of those that identify as Democrats and those that identify as Republicans (strongly or not strongly).

domains. Our calculations exclude independents (which have been increasing over this time period), but we find similar patterns when we include independents who lean towards either Democrats or Republicans.

Table 3 shows that the party gap in attitudes has grown significantly in every domain for both men and women. For example, in the 1970s, the party gap in the abortion attitude index for both sexes hovered around 2 percentage points, while in the 2010s the party gap grew to 21 and 14 percentage points for women and men, respectively. The party gap in attitudes towards sexuality similarly increased more for women. On both of these issues, a substantial 22 percentage point chasm has opened across Democratic and Republican women, compared with a 15 percentage point gap for men. On other issues, women and men have essentially converged to the same party gap, which now stands at two to three times the level of the 1970s.

Particularly striking are the 20 and 30 percentage point party gaps in the criminal justice and progressive government indices, respectively.

It thus appears that the profile of the Democratic and Republican voter, regardless of sex, is quite distinct from the past. This provides suggestive evidence in line with the hypothesis in Gillion, Ladd, and Meredith (2018) that changes in party sorting across sexes can reconcile the trends in the partisan sex gap and preferences that we observe.

Drivers of Sex Gaps and the Growing Political Influence of Women

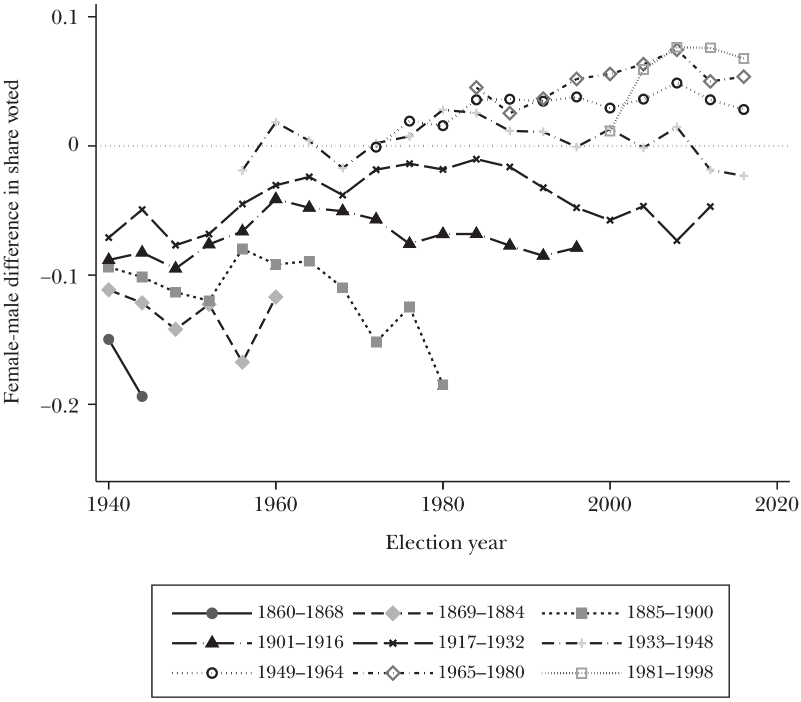
The findings thus far suggest that women are a much stronger political force today than they were immediately after adoption of the Nineteenth Amendment. Relative to men, women are now more likely to vote and more likely to identify as Democrats. What forces have generated the large observed changes in relative female mobilization and partisanship? Stronger partisans show up more reliably at the polls (Gerber, Huber, and Washington 2010). But this need not mean that the trends in the sex gaps in voter turnout and party identification shown in Figures 2 and 3 have the same root causes.

Cohort and Time Effects

For an exploratory analysis, we categorize potential forces changing the sex gap into two mutually exclusive groups: *cohort effects* that are constant across the life span of a given cohort, defined as a group of individuals born at roughly the same time, and *time effects* that affect individuals of all ages observed at the same time in the same way. With regard to cohort effects, experiences in early life may socialize women and men differently into (or away from) voting, and perhaps a later-life experience at the polls (or elsewhere) reinforces this early-life socialization (Gerber, Green, and Shachar 2003; Coppock and Green 2016; Fujiwara, Meng, and Vogl 2016). With regard to time effects, changes in society or in the policy stances that define parties may attract women of all ages.

Figures 5 and 6 show time trends in the sex gap in voter turnout and Democratic Party identification, respectively, separately by cohort. We group birth years into nine roughly 16-year cohorts that have a large degree of overlap with generations studied by demographers. These include individuals who would have been children when the Nineteenth Amendment was ratified, born between 1901 and 1916 (roughly the first half of the "Greatest Generation"), and individuals who would have been the youngest eligible voters in 1920, born between 1885 and 1900 (the "Lost Generation"). We also include two earlier cohorts comprised of individuals who would have been exposed to the Nineteenth Amendment at midlife (born between 1869 and 1884) or older (born between 1860 and 1868). Generations since the Nineteenth Amendment include individuals born between 1917 and 1933 (the second half of the Greatest Generation), 1933 and 1948 (the Silent Generation), 1949 and 1964 (Baby Boomers), 1965 and 1980 (Generation X), and 1981 and 1998 (Millennials).

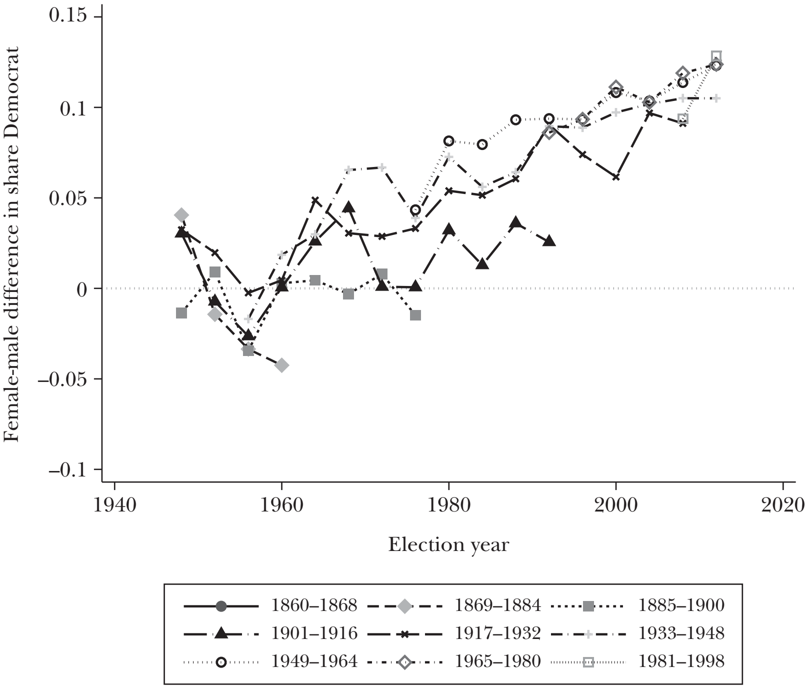
Figure 5
The Sex Gap in Voter Turnout by Cohort: Pooled Survey Data, 1940–2016



Source: Survey data pool the Gallup, November CPS, General Social Survey, and American National Election Studies series described in the notes to Figure 1.
Note: Statistics are weighted by survey-provided weights (for the CPS, General Social Survey, and American National Election Studies) or author-constructed weights (for Gallup), with all weights renormed to average to one within each survey-year. The figure plots the difference in estimated voter turnout rates between women and men by year and cohort. We omit cells based on small sample sizes (<150 observations per sex) to reduce noise.

The patterns of Figures 5 and 6 suggest, consistent with our earlier discussion, that cohort effects have been more important for the evolution of the sex gap in voter turnout and time effects more important for the evolution of the sex gap in party identification. While the cohort-specific time trends shown in Figure 5 are not literally flat (which is what a completely pure cohort effect would look like), younger generations typically show more positive sex gaps in turnout in every year. In addition, differences in the sex gap are on average greater across earlier cohorts than later ones, which corresponds with how the sex gap in turnout shrinks at a faster pace over earlier years of our sample period. In earlier research, Prior (2010) also shows that political mobilization is remarkably stable over the life cycle, using individual panel data. Also consistent with these findings, Firebaugh and Chen (1995) show that there is an especially large sex gap in voter turnout for the earliest cohort observable in data from the American National Election Studies.

Figure 6
Sex Gap in Democratic Party Identification by Cohort: Pooled Survey Data, 1948–2016



Source: Survey data pool the Gallup, General Social Survey, and American National Election Studies series described in the notes to Figure 3.
Note: Statistics are weighted by survey-provided weights (for the General Social Survey and American National Election Studies) or author-constructed weights (for Gallup), with all weights re-normed to average to one within each survey-year. The figure plots the difference in estimated rates of identification with the Democratic Party between women and men by year and cohort. We omit cells based on small sample sizes (<150 observations per sex) to reduce noise; for this reason, no observations from the earliest cohort are shown.

In contrast, the pattern in Figure 6 is better interpreted as a time-effect pattern, in which the sex gap in Democratic Party identification rises over time within each cohort. The result is smaller cross-cohort differences in the sex gap at a given point in time and a strong common upward trajectory. A regression analysis of data collapsed to the cohort-by-election year-by-state level confirms that cohort effects essentially completely explain the time trend in the sex gap in voter turnout but explain little of the time trend in the sex gap in Democratic Party identification.[11]

[11] The geographic unit to which we collapse is actually the single states and groups of states (27 total) identified in the 1976 November CPS. We omit the General Social Survey from this portion of the analysis, due to lack of information on state of residence in the public-use data. We regressed sex-by-cohort-by-state

Can the Cohort Effects Be Explained?

The relative contributions of cohort and time effects for these outcomes map to different sets of potential causal mechanisms for the evolution of sex gaps in voting behavior. Above, we presented evidence consistent with the inter-decadal growth in relative Democratic Party identification among women that cannot be explained by generational replacement: women of all ages have moved toward the Democrats as the two major parties have become increasingly divided on issues that women tend to care about. On the other hand, the growth in women's relative turnout appears to be largely explained by generational replacement. In this section, we consider the relevance of several cohort-varying factors for these findings.

Because our data are stratified by state, we can assess the explanatory power of both observed and unobserved factors that vary across cohorts. This is important, because some factors potentially contributing to cohort effects, like the "norms against [women's] political engagement" (Corder and Wolbrecht 2016, 14), will be difficult to quantify. On the other hand, other potentially important cohort-specific factors can be measured. We initially focused on educational attainment, employment rates, and divorce because they have been identified as important determinants of political behavior in prior research.[12] But because our central finding was that changes in the sex gap in turnout across cohorts tracked gains in education, we focus on the attainment results here. Even so, our findings should not be interpreted causally. We have used the extant variation, not exogenous variation, in educational attainment across cohorts and states, which may be correlated with other, unobserved state-by-cohort factors.

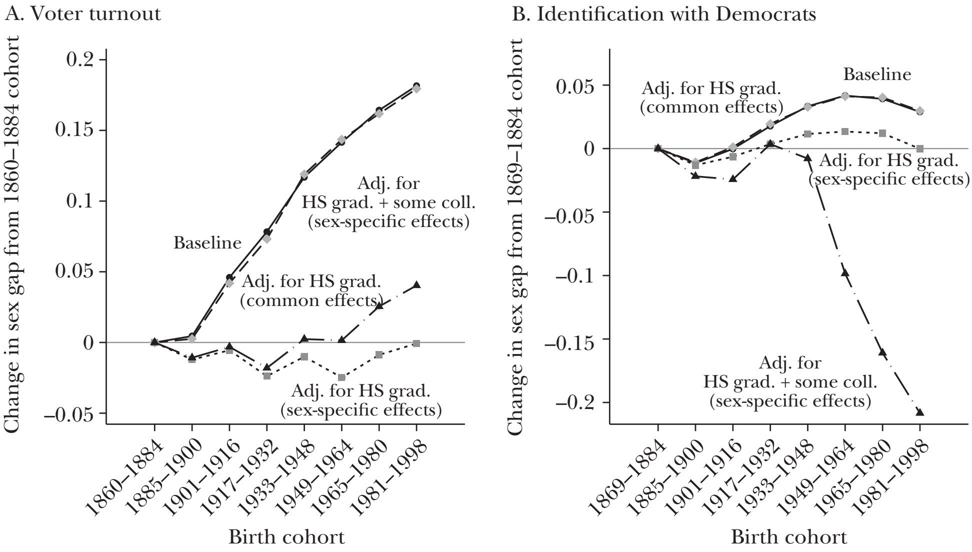
Baseline cohort effects in the sex gap in voter turnout, represented as an across-cohort change relative to the earliest two cohorts combined, are shown with the solid line in the first panel of Figure 7. For all cohorts beyond the first, these changes are significant both statistically and in magnitude. For example, the sex gap in turnout has been about 14 percentage points more favorable to Baby Boomer women than it was to women born roughly a century before.

To examine the role of rising levels of education in these patterns, we begin by introducing state-by-cohort-by-sex controls for high school completion as of age 25, estimated from the census and American Community Survey. Because cross-cohort gains in high school were similar by sex, we do not expect this to have much explanatory power, and in fact, we see little impact on our estimated cohort effects in panel A of Figure 7. Next, we allow for changes in high school completion to have different effects by sex, thus allowing for the realistic possibility that education

group-by-year outcomes on decade indicators interacted with a female indicator in a model including fixed effects for sex, decade, state group, and sex-by-state group. Online Appendix Figure 7 shows what happens to the coefficients on the decade-by-female interaction terms with the addition of cohort and cohort-by-sex fixed effects to this model. For details and further explication of the regression, see the online Appendix available with this paper at the *JEP* website.

[12] For example, divorce and economic vulnerability have been linked to the rising sex gap in Democratic Party identification in the United States (Edlund and Pande 2002; Box-Steffensmeier, De Boef, and Lin 2004).

Figure 7
Educational Attainment and the Cohort Effects in Voting Sex Gaps



Note: The figures plot coefficients from regressions where the dependent variable is voter turnout (panel A) or the rate of Democratic Party identification (panel B) at the election year × cohort × sex × state group level. All regressions are weighted by the number of observations used to construct the dependent variable. The solid line ("baseline") plots the change in the sex gap in the outcome from the initial birth cohort (1860–1884 in panel A and 1869-1884 in panel B), or coefficients on the interaction between a female dummy and cohort indicators from a version of the regression described in the online data Appendix that also includes cohort fixed effects. Other lines show what would have happened to the sex gap in voter turnout across cohorts holding constant high school graduation and some college completion. Throughout, the sex gap is the female-male difference in the outcome.

could have a different impact on women's political participation. Because the positive association between high school completion and lifetime political mobilization is significantly stronger for women than for men, much of the unobserved cohort effects in the sex gap in turnout fades away.[13] Although women born in the mid-1950s and later have been more likely to attend and complete college than their male counterparts (Goldin, Katz, and Kuziemko 2006), adjusting for college attendance by age 25 offers little additional explanatory power, given the comparatively weak association between college attendance and turnout in our data.[14]

Thus, the rise in educational attainment—or, perhaps, other correlated outcomes—appears to explain the changes in the sex gap in voter turnout across

[13] The association in our data for women is similar to that found by Milligan, Moretti, and Oreopoulos (2004) for the population overall exploiting variation in completion from compulsory schooling and child labor laws early in the twentieth century.

[14] We added each variable directly and interacted with a female indicator to the model outlined in footnote 11. For detailed regression findings, see online Appendix Table 3.

cohorts; adding sex-specific effects of high school completion in particular to the model lowers and renders statistically insignificant the contribution of unobserved cohort-specific factors. The change is especially noticeable for cohorts born in the first half of the twentieth century, for whom changes in high school completion were particularly dramatic even if similar across sex. The sizable sex difference in the association between high school completion and turnout in our data could be explained by the particular role that education plays in women's lives, such as through reductions in fertility, or through other factors coinciding with the high school movement—such as advancement of social norms—that could have pushed women's voter participation up more than men's.

The solid line in the other panel of Figure 7 shows the baseline, essentially negligible cohort effects in the sex gap in Democratic Party identification, consistent with Figure 6. Though not associated with turnout in our data, college attendance is positively associated with increases in Democratic partisanship and more so for women. Because of the divergence of women's college attendance rates from men's across recent cohorts, holding constant college attendance thus *generates* some unexplained cohort effects. As shown in the figure, in the absence of rising college attendance, Baby Boomer, Gen-X, and Millennial women would have actually been *less* likely to identify with the Democratic Party.

Taken as a whole, these descriptive findings suggest that each successive generation of women has been more politically mobilized than her predecessors, with educational attainment playing an important role. However, this is only a descriptive exercise that would be useful to revisit.

Conclusion

The female voter has come a long way since the passage of the first suffrage laws at the turn of the century and since the passage of the Nineteenth Amendment in 1920 extended the franchise (at least in principle) to women nationwide. We trace the evolution of the sex gap in voter turnout and partisanship over the last 80 years using a novel dataset of voter surveys. We find that women closed a 10 percentage point gap in voter turnout over the 40 years from 1940 to 1980 and over the next 40 years from 1980 to present gained more than a 4 percentage point advantage in turnout over men. Additionally, while women and men had similar patterns of party support in 1940, over the last half-century, a 12 percentage point sex gap has emerged in the probability of women and men identifying with the Democratic Party.

What accounts for these changes? We argue that the relative rise in women's turnout is largely explained by the replacement of older, low-participation cohorts with younger, high-participation cohorts. Descriptively, we find that these cohort effects are associated with women's differential response to increasing rates of high school graduation, with less explanatory power for rising rates of college attendance. In contrast, the rise in women's support for Democrats appears to have been common to all cohorts. At least since the 1970s, this seems to be best explained by the trend towards greater polarization of political parties, as we find little evidence of any change in the gap in policy preferences across men and women.

Many gaps remain in analyzing the causes and consequences of this century of political progress for women. First, what are the causal factors behind the large rise in women's voter turnout across cohorts? To the best of our knowledge, there is no research providing a credible analysis of the link between the increase in voter participation that we have documented and the significant advances made by women across cohorts—in educational attainment, economic opportunities, and access to contraceptive technology, to name a few—despite the fact that these changes appear to have occurred simultaneously. Our descriptive analyses suggest that rising education may have the most explanatory power, but a more rigorous design may yield different results. Second, in what ways have the rise in women's voter participation and greater identification with the Democratic Party affected modern policy outcomes? In addition to clarifying the process of political change for women, providing answers to these questions may also provide broader insights into the process of acquiring political capital for newly enfranchised groups.

■ *We thank Andrew Ahn and Catalina Garcia Valenzuela for outstanding research assistance. We also thank Christina Wolbrecht, Ebonya Washington, Dawn Teele, Dean Lacy, and Lisa Baldez for helpful comments on the first draft of this paper.*

References

American National Election Studies. 1948–2016. "American National Election Studies Time Series Cumulative Data File." http://www.electionstudies.org/ (accessed March 19, 2019).

Ansolabehere, Stephen, and Eitan Hersh. 2012. "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate." *Political Analysis* 20 (4): 437–59.

Bailey, Martha J. 2006. "More Power to the Pill: The Impact of Contraceptive Freedom on Women's Life Cycle Labor Supply." *Quarterly Journal of Economics* 121 (1): 289–320.

Berinsky, Adam J. 2006. "American Public Opinion in the 1930s and 1940s: The Analysis of Quota-Controlled Sample Survey Data." *Public Opinion Quarterly* 70 (4): 499–529.

Bernstein, Robert, Anita Chadha, and Robert Montjoy. 2001. "Overreporting Voting: Why It Happens and Why It Matters." *Public Opinion Quarterly* 65 (1): 22–44.

Besley, Timothy, and Stephen Coate. 1997. "An Economic Model of Representative Democracy." *Quarterly Journal of Economics* 112 (1): 85–114.

Bolzendahl, Catherine I., and Daniel J. Myers. 2004. "Feminist Attitudes and Support for Gender Equality: Opinion Change in Women and Men, 1974–1998." *Social Forces* 83 (2): 759–89.

Box-Steffensmeier, Janet M., Suzanna De Boef, and Tse-min Lin. 2004. "The Dynamics of the Partisan Gender Gap." *American Political Science Review* 98 (3): 515–28.

Carruthers, Celeste K., and Marianne H. Wanamaker. 2015. "Municipal Housekeeping: The Impact of

Women's Suffrage on Public Education." *Journal of Human Resources* 50 (4): 837–72.

Carsey, Thomas M., and Geoffrey C. Layman. 2006. "Changing Sides or Changing Minds? Party Identification and Policy Preferences in the American Electorate." *American Journal of Political Science* 50 (2): 464–77.

Cascio, Elizabeth U., and Ebonya Washington. 2014. "Valuing the Vote: The Redistribution of Voting Rights and State Funds following the Voting Rights Act of 1965." *Quarterly Journal of Economics* 129 (1): 379–433.

Cassese, Erin C., and Tiffany D. Barnes. 2018. "Reconciling Sexism and Women's Support for Republican Candidates: A Look at Gender, Class, and Whiteness in the 2012 and 2016 Presidential Races." *Political Behavior* 41 (3): 677–700.

Clark, April K. 2017. "Updating the Gender Gap(s): A Multilevel Approach to What Underpins Changing Cultural Attitudes." *Politics & Gender* 13 (1): 26–56.

Coppock, Alexander, and Donald P. Green. 2016. "Is Voting Habit Forming? New Evidence from Experiments and Regression Discontinuities." *American Journal of Political Science* 60 (4): 1044–62.

Corder, J. Kevin, and Christina Wolbrecht. 2016. *Counting Women's Ballots: Female Voters from Suffrage through the New Deal.* New York: Cambridge University Press.

Cox, Gary W., and Matthew D. McCubbins. 1986. "Electoral Politics as a Redistributive Game." *Journal of Politics* 48 (2): 370–89.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2): 448–74.

DiMaggio, Paul, John Evans, and Bethany Bryson. 1996. "Have American's Social Attitudes Become More Polarized?" *American Journal of Sociology* 102 (3): 690–755.

Downs, Anthony. 1957. "An Economic Theory of Political Action in a Democracy." *Journal of Political Economy* 65 (2): 135–50.

Duflo, Esther. 2012. "Women Empowerment and Economic Development." *Journal of Economic Literature* 50 (4): 1051–79.

Edlund, Lena, and Rohini Pande. 2002. "Why Have Women Become Left-Wing? The Political Gender Gap and the Decline in Marriage." *Quarterly Journal of Economics* 117 (3): 917–61.

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

Fernández, Raquel, Sahar Parsa, and Martina Viarengo. 2019. "Coming Out in America: AIDS, Politics, and Cultural Change." NBER Working Paper 25697.

Filer, John E., Lawrence W. Kenny, and Rebecca B. Morton. 1991. "Voting Laws, Educational Policies, and Minority Turnout." *Journal of Law and Economics* 34 (2): 371–93.

Firebaugh, Glenn, and Kevin Chen. 1995. "Vote Turnout of Nineteenth Amendment Women: The Enduring Effect of Disenfranchisement." *American Journal of Sociology* 100 (4): 972–96.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren. 2018. "Integrated Public Use Microdata Series, Current Population Survey: Version 6.0." IPUMS, Minneapolis, MN. https://doi.org/10.18128/D030.V6.0 (accessed January 21, 2020).

Fogli, Alessandra, and Laura Veldkamp. 2011. "Nature or Nurture? Learning and the Geography of Female Labor Force Participation." *Econometrica* 79 (4): 1103–38.

Fujiwara, Thomas, Kyle Meng, and Tom Vogl. 2016. "Habit Formation in Voting: Evidence from Rainy Elections." *American Economic Journal: Applied Economics* 8 (4): 160–88.

Gerber, Alan S., Donald P. Green, and Ron Shachar. 2003. "Voting May Be Habit-Forming: Evidence from a Randomized Field Experiment." *American Journal of Political Science* 47 (3): 540–50.

Gerber, Alan S., Gregory A. Huber, and Ebonya Washington. 2010. "Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment." *American Political Science Review* 104 (4): 720–44.

Gerring, John. 1998. *Party Ideologies in America, 1828–1996.* Cambridge: Cambridge University Press.

Gillion, Daniel Q., Jonathan M. Ladd, and Marc Meredith. 2018. "Party Polarization, Ideological Sorting and the Emergence of the US Partisan Gender Gap." *British Journal of Political Science* 1–27.

Goldin, Claudia. 1990. *Understanding the Gender Gap: An Economic History of American Women.* New York: Oxford University Press.

Goldin, Claudia, Lawrence F. Katz, and Ilyana Kuziemko. 2006. "The Homecoming of American College Women: The Reversal of the College Gender Gap." *Journal of Economic Perspectives* 20 (4): 133–56.

Inglehart, Ronald, and Pippa Norris. 2000. "The Developmental Theory of the Gender Gap: Women's and Men's Voting Behavior in Global Perspective." *International Political Science Review* 21 (4): 441–63.

Kaufmann, Karen M., and John R. Petrocik. 1999. "The Changing Politics of American Men: Understanding the Sources of the Gender Gap." *American Journal of Political Science* 43 (3): 864–87.

Keyssar, Alexander. 2009. *The Right to Vote: The Contested History of Democracy in the United States* [Revised Edition]. New York: Basic Books.

Kleppner, Paul. 1982. "Were Women to Blame? Female Suffrage and Voter Turnout." *Journal of Interdisciplinary History* 12 (4): 621–43.

Kose, Esra, Elira Kuka, and Na'ama Shenhav. 2019. "Who Benefited from Women's Suffrage?" http://www.elirakuka.com/uploads/1/0/0/6/10064254/kks_sep19.pdf.

Kuziemko, Ilyana, and Ebonya Washington. 2018. "Why Did the Democrats Lose the South? Bringing New Data to an Old Debate." *American Economic Review* 108 (10): 2830–67.

Lawless, Jennifer L., and Richard L Fox. 2013. *Girls Just Wanna Not Run: The Gender Gap in Young Americans' Political Ambition.* Washington, DC: Women & Politics Institute.

Lawless, Jennifer L., and Kathryn Pearson. 2008. "The Primary Reason for Women's Underrepresentation? Reevaluating the Conventional Wisdom." *Journal of Politics* 70 (1): 67–82.

Layman, Geoffrey C., and Thomas M. Carsey. 2002. "Party Polarization and 'Conflict Extension' in the American Electorate." *American Journal of Political Science* 46 (4): 786–802.

Lott, John R., Jr., and Lawrence W. Kenny. 1999. "Did Women's Suffrage Change the Size and Scope of Government?" *Journal of Political Economy* 107 (6): 1163–98.

Miller, Grant. 2008. "Women's Suffrage, Political Responsiveness, and Child Survival in American History." *Quarterly Journal of Economics* 123 (3): 1287–1327.

Milligan, Kevin, Enrico Moretti, and Philip Oreopoulos. 2004. "Does Education Improve Citizenship? Evidence from the United States and the United Kingdom." *Journal of Public Economics* 88 (9–10): 1667–95.

Moehling, Carolyn M., and Melissa A. Thomasson. 2012. "The Political Economy of Saving Mothers and Babies: The Politics of State Participation in the Sheppard-Towner Program." *Journal of Economic History* 72 (1): 75–103.

Morgan-Collins, Mona. Forthcoming. "The Electoral Impact of Newly Enfranchised Groups: The Case of Women's Suffrage in the United States." *Journal of Politics.*

Osborne, Martin J., and Al Slivinski. 1996. "A Model of Political Competition with Citizen-Candidates." *Quarterly Journal of Economics* 111 (1): 65–96.

Prior, Markus. 2010. "You've Either Got It or You Don't? The Stability of Political Interest over the Life Cycle." *Journal of Politics* 72 (3): 747–66.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2019. "IPUMS USA: Version 9.0." IPUMS, Minneapolis, MN. https://doi.org/10.18128/D010.V9.0 (accessed January 9, 2020).

Shapiro, Robert Y., and Harpreet Mahajan. 1986. "Gender Differences in Policy Preferences: A Summary of Trends from the 1960s to the 1980s." *Public Opinion Quarterly* 50 (1): 42–61.

Smith, Tom W., Michael Davern, Jeremy Freese, and Stephen L. Morgan. 2019. *General Social Surveys, 1972–2018.* Chicago: National Opinion Research Center.

Teele, Dawn Langan. 2018. *Forging the Franchise: The Political Origins of the Women's Vote.* Princeton: Princeton University Press.

US Census Bureau. 1992. "Current Population Survey: Voter Supplement File, 1972." Inter-University Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/ICPSR00060.v1 (accessed March 25, 2019).

Valelly, Richard M. 2004. *The Two Reconstructions: The Struggle for Black Enfranchisement.* Chicago: University of Chicago Press.

Wasserman, Melanie. 2018. "Gender Differences in Politician Persistence." https://pdfs.semanticscholar.org/523a/964eb59f12b1e6c4db51ce1203fbdb5164ad.pdf.

Wolbrecht, Christina, and J. Kevin Corder. Forthcoming. *A Century of Votes for Women: American Elections since Suffrage.* New York: Cambridge University Press.

Sociological Perspectives on Racial Discrimination

Mario L. Small and Devah Pager

Race discrimination has long been a focus of research in sociology. Sociologists generally understand racial discrimination as differential treatment on the basis of race that may or may not result from prejudice or animus and may or may not be intentional in nature. This understanding of discrimination has produced a rich and diverse research body of work with both similarities to and differences from traditional research on the topic in economics.

Much of the sociological research on race discrimination will be familiar to economists. As in economics, much of it has focused on discrimination in employment, housing, and credit markets (for example, Pager and Shepherd 2008; Fernandez-Mateo 2009; Gaddis 2015). As in economics, some sociological studies are based on observational data, with statistical models where the outcome of interest, such as wages or employment, is regressed on a race indicator variable and on other variables that could account for the effect of race. As economists have, sociologists have noted that this "residual race gap" approach may be vulnerable to the consequences of unobserved heterogeneity (Cancio, Evans, and Maume 1996; Farkas and Vicknair 1996). As in economics, many studies in sociology are based on field experiments, such as audit or correspondence studies (for reviews by sociologists, see Pager and Shepherd 2008; Quillian et al. 2017). Sociologists

■ *Mario L. Small is Grafstein Family Professor of Sociology, Harvard University, Cambridge, Massachusetts. Devah Pager was Professor of Sociology and Public Policy, Harvard University, Cambridge, Massachusetts, at the time of her death on November 2, 2018. Small's email address is mariosmall@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.34.2.49.

have also noted the problems with experimental studies, including that they are often based on data not representative of the national population and that they may not indicate how much discrimination there is in actual markets. And as in economics, quite a few researchers have used the study of discrimination to examine larger methodological questions (Pager 2007).

However, much of the sociological work on discrimination is more distinctive to the discipline. Part of the reason is the discipline's methodological heterogeneity. Sociology includes researchers who analyze large-*n* observational data (as demographers do); who observe organizations, neighborhoods, and villages ethnographically (as anthropologists do); who conduct either laboratory or field experiments (as psychologists do); and who examine historical documents through close readings (as historians do) (Small 2011). This heterogeneity shapes how researchers think about and assess evidence, such that sociologists differ among themselves in what they take to be most important in an empirical study. Many sociologists, influenced by the causal revolution across the social sciences, prioritize the ability to make convincing causal claims (Morgan and Winship 2015). Other sociologists are comfortable analyzing data that do not permit bulletproof causal claims when the data have other critical advantages, such as allowing generalizability about a national population or providing access to a process otherwise difficult to observe (for example, Hultin and Szulkin 1999; Turco 2010). To their point, some of the sociological research most impactful across the social sciences, including economics, has been associational in nature, such as Wilson's (1987, 1996) highly insightful work on how joblessness affects the social fabric in urban neighborhoods.

In fact, many questions important to understanding discrimination are purely descriptive, such as how much discrimination people anticipate in job or real estate markets; how employers perceive black job applicants; or how landlords, bankers, or others gatekeepers understand their behavior (for example, Pager and Karafin 2009; Light, Roscigno, and Kalev 2011; Kang et al. 2016). Thus, the sociology of discrimination involves both descriptive and causal research, and our discussion does not prioritize one over the other. Being cognizant of this heterogeneity in sociology with respect to methods and perspectives will be useful in assessing our discussion.

In what follows, we offer six propositions from the sociology of racial discrimination that we believe are worth noting by economists. We do not consider these to be the six ideas most central to the sociology of discrimination, or even the six most often studied. Instead, it is a list of propositions we consider to be particularly sociological in perspective, notably different from traditional research in the economics of discrimination, and worth greater attention by researchers in both disciplines. Some of these propositions are reflected in recent research in economics and may represent a bridge between the disciplines. Readers seeking comprehensive reviews of the sociology of discrimination might begin with National Research Council (2004), Lucas (2008), and Pager and Shepherd (2008).

Taste and Statistical Discrimination Do Not Capture All Reasons behind Differential Treatment by Race

Research on discrimination in economics has traditionally adopted one of two perspectives. One is the "taste for discrimination" perspective, which posits that people discriminate when they are willing to pay a price of some kind to reduce their association or interaction with a given group (Becker 1971, 14). Becker, who developed this perspective in the 1950s, examined the dynamics likely to evolve as a result of discrimination among employers, other employees, and customers; the many researchers who followed applied the perspective to a broad range of actors, including realtors, employers, and bankers.

The other common approach in economics is the "statistical discrimination" perspective, which posits that employers, facing limited information about a given potential employee, use group characteristics to make inferences about those of the individual (Arrow 1972a, 1972b; Phelps 1972). As Phelps (1972, 659) put it: "[T]he employer who seeks to maximize expected profit will discriminate against blacks or women if he believes them to be less qualified, reliable, long-term, etc., on the average than whites and men, respectively, and if the cost of gaining information about the individual applicants is excessive. Skin color or sex is taken as a proxy for relevant data not sampled." This perspective does not require employers to hold racial animus. Though the two perspectives differ in many important ways, the core difference for our purposes is on why people discriminate: either because they are hoping to reduce contact with a member of a different group, or because they are behaving as would a rational actor who is either unable or unwilling to acquire additional information.

For many years, economists tended to adopt either one or the other perspective, and as Guryan and Charles (2013, F417) report, economists recently have "returned to the question of whether taste-based or statistical discrimination is a more appropriate description of the phenomenon."[1] This focus represents an important contrast to sociology. Many sociologists have studied racial prejudice (for reviews, see Reskin 2000; Bobo et al. 2012). But very few sociologists have adopted a statistical discrimination perspective and probably fewer have attempted to determine whether what looks like prejudice-based discrimination may in fact be statistical discrimination. Since sociology has never had a de facto, agreed-upon view of decision-making in which actors are rational, demonstrating that an act of discrimination can be shown to be accounted for by reasonable guesses based on group characteristics given available information would not offer much *sociological* insight (see also Pager and Karafin 2009).

[1] The research on taste-based and statistical discrimination is naturally far more sophisticated than the views we sketch here. For example, a researcher adopting the statistical discrimination perspective may assume that differences in groups' characteristics observed by employers partly reflect prior discrimination. This idea would not be inconsistent with some of the forms of discrimination we discuss below.

In addition, many sociologists would offer a critique applying to both perspectives—both assume, or appear to assume, that actors are making decisions to discriminate deliberately. Long-standing research in psychology, specifically on implicit bias, would question this assumption. In fact, a group of economists informed by this work has recently proposed a third perspective. "Under both [conventional] models," they write, "individuals *consciously* discriminate, either for a variety of personal reasons or because group membership provides information about a relevant characteristic, such as productivity. Motivated by a growing body of psychological evidence, we put forward a third interpretation: implicit discrimination. Sometimes, we argue, discrimination may be *unintentional* and outside of the discriminator's awareness" (Bertrand, Chugh, and Mullainathan 2005, 94). While the implicit bias perspective shares with Becker's taste model the idea that discrimination happens when individuals are racially prejudiced, it does not view individuals as rationally balancing their taste for discrimination against the price they pay for exercising that taste.

The questions of intention in and consciousness about decision-making are largely a domain of psychology, rather than sociology, and we refer readers to Bertrand, Chugh, and Mullainathan (2005) for a review of that work in the context of economic research. Nonetheless, the core critique is worth noting. Discrimination from implicit biases does not require people to be making the conscious decisions that both conventional approaches would seem to presume.

We would go further. Though the perspectives of taste for discrimination, statistical discrimination, and implicit bias make different assumptions about why discrimination happens, they all agree on a core issue: for discrimination to happen, an individual must decide to treat people of different backgrounds differently (even if the decision is driven by an unconscious bias). As a result, they miss what sociologists and others have called "institutional discrimination," "structural discrimination," and "institutional racism," which are all terms used to refer to the idea that something other than individuals may discriminate by race (Feagin and Eckberg 1980; Massey and Denton 1993; Oliver and Shapiro 2006; Pager and Shepherd 2008; Reskin 2012). Unfortunately, these terms are not used consistently across the social sciences; moreover, they are often used even more ambiguously among lay writers and commentators. Nevertheless, a substantial body of evidence suggests that limiting the study of discrimination to the actions of potentially prejudiced individuals dramatically understates the extent to which people experience discrimination; understates the extent to which discrimination may account for social inequality; and understates the extent to which discrimination may play a role in markets for labor, credit, and housing, as well as in other contexts.

In this essay, we define "institutional discrimination" as differential treatment by race that is either perpetrated by organizations or codified into law. Because discrimination may be caused by organizational rules or by people following the law, it need not result from personal prejudice, from rational guesses on the basis of group characteristics, or from implicit racism. Institutional discrimination can take different forms, and we cannot hope to cover either all of them or their relation to institutional sociology in these pages (for discussions, see National Research Council 2004; also Powell and DiMaggio 1991; Scott 2013). But as we will take pains to make clear, understanding several of these forms requires a more expansive view of discrimination.

Organizations Can Discriminate Irrespective of the Intentions of Their Members

We define an organization as a loosely coupled set of people and institutional practices formally organized around a global purpose (Small 2009; Scott 2013). Examples include banks, universities, churches, childcare centers, real estate agencies, unions, and country clubs. Note that an organization is not just its people but also its institutional practices. A practice can be "institutional" either normatively or cognitively, in the sense that it can be shaped by either norms or "cognitions" (Small 2009; Small, Harding, and Lamont 2010).

A norm is a formal or informal expectation of behavior that people in the organization feel compelled to follow. For example, a university's requirement that tenure cases include external evaluations is a formal norm (a rule); its expectation that faculty be polite to their students, an informal one. In both cases, people generally feel some compulsion to follow the norm (though they may of course choose to violate it). A cognition—sometimes referred to as "frames" or "cognitive understandings" in sociology—is not a mandate of any kind; instead, it is a way of understanding one's predicament in the organization. For example, whether students believe the economics major is prestigious is a cognitive understanding, not a formal or informal mandate. Institutional practices, whether normative or cognitive, are generally understood by members of the organization. However, an organization's institutional practices generally do not depend on any particular individual, in the sense that they may be stable even as people enter and leave the organization. This independence is part of the reason sociologists of organizations take pains to distinguish individuals from the normative expectations or cognitive understandings that shape their behavior.

Organizations can discriminate when they have instituted practices, formally or informally, that treat people of different races differently, regardless of whether the practices were driven by prejudice and regardless of whether the managers, directors, or employees following the norms are themselves racially prejudiced. One example is a hiring norm common among US employers. Researchers have documented that many companies fill job vacancies through referral networks, wherein employees are formally or informally asked to recommend candidates for positions (for example, Mouw 2002; Waldinger and Lichter 2003; also Arrow 1998). Sociologists have also documented repeatedly that social networks are racially homophilous, in the sense that people tend strongly to have friends of their own race (McPherson, Smith-Lovin, and Cook 2001). If so, then a racially homogeneous organization composed of non-prejudiced actors that has instituted referral-based employment will tend to hire relatively few people of a different race. In this sense, the organization is institutionally discriminating, because people of a different race have little chance of landing a job there.

Of course, in practice, this process is far more complicated; for example, empirical research suggests that whether referral-based hiring produces this result depends on many other factors, including how racially diverse the local region is and how jobholders decide to share information about openings (Fernandez and Fernandez-Mateo 2006; Rubineau and Fernandez 2013). But the process makes clear that an organization may discriminate independent of a manager's explicit decision to hire candidates of a given race (whether statistically or from prejudice).

Many organizational processes with discriminatory consequences have a similar form: an institutional practice that is in theory race-neutral affects racial minorities because it is applied in a context with a preexisting racial difference, gradient, or level of segregation.[2] Consider downsizing and layoffs. Many companies in the process of downsizing seek to reduce exposure to liability by instituting formal processes that are presumably designed to avoid bias, such as laying off managers based on their tenure with the organization or on the importance of their position. But the managerial ranks of many organizations have only recently, in historical terms, included large numbers of minorities and women. Furthermore, in many companies, racial minorities are less likely to hold the more important managerial positions (Elliott and Smith 2004)—for example, racial minorities are more likely to be managers in benefits or community outreach units than in critical operations.

In this context, a formal rule that layoffs are based on years of employment or on the importance of the position will reduce the number of minority and women managers. It is not surprising that a national study of 327 establishments that downsized between 1971 and 2002 found that downsizing reduced the diversity of the firm's managers—female and minority managers tended to be laid off first. But what is perhaps more surprising is that those companies whose layoffs were based formally on tenure or position saw a greater decline in the diversity of their managers; net of establishment characteristics such as size, personnel structures, unionization, programs targeting minorities for management, and many others; and of industry characteristics such as racial composition of industry and state labor force, proportion of government contractors, and others (Kalev 2014). In contrast, those companies whose layoffs were based formally on individual performance evaluations did not see greater declines in managerial diversity (Kalev 2014). Patterns of this kind help to explain why sociologists have paid increasing attention to institutional practices when studying the racial composition of staff in organizations (Dobbin 2009; Kalev 2014; Dobbin, Schrage, and Kalev 2015; also Bielby 2000).

[2] Readers may note a similarity between our discussion of institutional discrimination by organizations and "disparate impact" discrimination in US law, wherein a violation of Title VII of the Civil Rights Act of 1964 may be found if employers use a practice that is race-neutral on its face but disproportionally affects a racial group adversely. Our discussion, entirely sociological in intent, is informed by organizational and institutional theory in sociology, not by legal scholarship. We leave to legal scholars whether and how our discussion relates to the scholarship on disparate impact.

Institutional practices are powerful. Much of their power stems from their stability or inertia, as illustrated by the fact that they regularly endure the departure of old leaders and the arrival of new ones. Consider a university that, as most probably do, requires the tenure process to include external evaluations by experts. A new university president is unlikely to want to do away with this practice; even if she tried, she would likely meet resistance; and if she succeeded, she would likely face lingering questions about her legitimacy as a leader. A rational university president would probably not bother spending time and energy in an attempt to change this rule, even if that new president believed external evaluations slowed the tenure process, were not reliable because of favoritism or other biases, or did not provide useful information not already contained in the publication record.

Not all practices are as firmly institutionalized in organizations as external review in universities. But many practices such as hiring based on referral, layoffs based on length of service, blanket background checks, and many others that potentially lead to a pattern of differential treatment by race are deeply institutionalized across organizations. These practices are long-established, taken for granted, and subject to inertia, and managers are unlikely to think of them as open to change. In fact, they routinely survive the 100 percent staff turnover that long-lasting organizations eventually experience. As a result, a research program focused only on the potential decisions of a contemporaneous manager or gatekeeper will likely miss a lot of what shapes the potentially discriminatory actions of organizations.

Historic Discrimination Has Contemporary Consequences (via Organizations)

Institutional factors can also matter via the contemporary consequences of past discrimination (Wilson 1978). Some kinds of discrimination in the past were so widespread that they resulted in major differences across racial groups whose consequences can still be detected. Furthermore, because many forms of discrimination in the past were institutional in nature, their consequences may still be observed in the practices of organizations or in the laws in place today (in a way, analogous to the case of external review letters in universities). As a result, even if all forms of discrimination, individual or institutional, taste-based or statistical, were to suddenly cease, there would still be multiple reasons to examine discrimination in the past to understand the present. This topic is complex and wide-ranging, and it includes a lot of research done not by sociologists but by historians (for example, Jackson 1985; Sugrue 1996; Hillier 2003). But two cases, discussed in this section and the next one, will illustrate its significance.

One well-documented case is the institutionalization of redlining in real estate. In this case, both changes in federal law and the organizations created to implement them, particularly the Home Owners Loan Corporation (HOLC) and the Federal Housing Administration (FHA), were important. (The discussion that follows is based on Jackson 1980; Massey and Denton 1993; and Hillier 2003, 2005; see also Crossney and Bartelt 2005.) The HOLC was created in 1933 to reduce foreclosures during the Great Depression. One of its most important creations was self-amortizing loans with uniform payments that extended to 20 years, rather than the typical shorter loans for which payments might still be due after their terms expired. This innovation would be essential for the later accessibility of homeownership to millions of Americans. The FHA was created in 1934 to help stabilize mortgage markets and to encourage home building so as to expand jobs in construction (Jackson 1980). The FHA insured home loans issued by banks, and at the majority of the assessed value of the property, so that the down payments homeowners needed to produce were reduced from about half to about 10 percent of the value. The FHA adopted the self-amortizing loans approach and extended the amortization rate by another five to ten years. The institutionalization of these new mortgage practices dramatically expanded homeownership among Americans.

However, the Home Owners Loan Corporation and Federal Housing Administration were also responsible for the spread of redlining. As part of its evaluation of whom to help, the HOLC created a formalized appraisal system, which included the characteristics of the neighborhood in which the property was located. Neighborhoods were graded from A to D, and those with the bottom two grades or rankings were deemed too risky for investment. Color-coded maps helped assess neighborhoods easily, and the riskiest (grade D) neighborhoods were marked in red. These assessments openly examined a neighborhood's racial characteristics, as "% Negro" was one of the variables standard HOLC forms required field assessors to record (for example, Aaronson, Hartley, and Mazumder 2019, 53; Norris and Baek 2016, 43). Redlined neighborhoods invariably had a high proportion of African-Americans. Similarly, an absence of African-Americans dramatically helped scores. For example, a 1940 appraisal of neighborhoods in St. Louis by the Home Owners Loan Corporation gave its highest rating, A, to Ladue, an area at the time largely undeveloped, described as "occupied by 'capitalists and other wealthy families'" and as a place that was "not the home of 'a single foreigner or Negro'" (Jackson 1980, 425). In fact, among the primary considerations for designating a neighborhood's stability were, explicitly, its "protection from adverse influences," "infiltration of inharmonious racial or nationality groups," and presence of an "undesirable population" (as quoted in Hillier 2003, 403; Hillier 2005, 217).

The Federal Housing Administration required a mandatory appraisal of the neighborhood as part of its guarantee and, through this and other means, disseminated redlining as a lending practice. Building on the systems and maps of the Home Owners Loan Corporation, the FHA fielded its own surveys, created some of its own maps, and developed its own analyses. It also disseminated its ideas about neighborhood risk through its widely distributed *Underwriting Manual* (Hillier 2003, 403). Consistent with racial attitudes of the time, the FHA, as historian Kenneth Jackson writes, "was extraordinarily concerned with 'inharmonious racial or nationality groups.'" Homeowners and financial institutions alike feared that an entire area could lose its investment value if rigid white-black separation was not maintained. The *Underwriting Manual* bluntly warned, "If a neighborhood is to retain stability, it is necessary that properties shall continue to be occupied by the same social and racial classes," and openly recommended "enforced zoning, subdivision regulations, and suitable restrictive covenants . . ." (Jackson 1980, 436). The restriction described in these covenants was the exclusion of Jews, blacks, and others from neighborhoods through formal agreements among neighbors (Massey and Denton 1993). In short, the FHA was informing lenders explicitly that its insurance program opposed racial integration. Just as the redlined maps made it difficult for blacks to receive favorable loans in predominantly black neighborhoods, the restrictions on racial integration made it hard for them to move to white or racially mixed ones.

Of course, these racial attitudes—both the negative perception of African-Americans and the preference for racially segregated neighborhoods—were not invented by the Home Owners Loan Corporation or the Federal Housing Administration; such attitudes were common at the time, including among lenders, realtors, and white home purchasers. What the HOLC and then the FHA did was *institutionalize* these attitudes through several specific mechanisms: creating a formal system of risk assessment, requiring the assessment in order to insure loans, adding neighborhood characteristics to the assessment, tying lower neighborhood grades to the presence of African-Americans, ensuring less favorable rates to lower neighborhood grades, discouraging racial integration, and spreading this particular bundle of cognitive understandings and normative expectations to lenders throughout the country via formal underwriting guidelines.

Moreover, by insuring millions of homes and multi-family projects worth billions of dollars, the federal agencies held enormous power over how lenders did their work. Banks that wanted to participate in the federal largesse would follow the guidelines of the Federal Housing Administration, resulting, inevitably, in fewer and less favorable loans to African-Americans—via the same mechanism through which the federal government was making home purchasing and wealth accumulation easier for others. Black purchasers hoping to secure mortgages from participating lenders would either have difficulty or face less favorable rates, because predominantly black neighborhoods were invariably rated D or redlined, and their move to predominantly white neighborhoods would clash with early restrictions against "inharmonious racial groups."

The assignment of lower scores to predominantly black neighborhoods by the Home Owners Loan Corporation and the Federal Housing Administration may well have reflected the realities of the market at that time. Whites were a substantial majority of home purchasers, and we know that whites at the time expressed extremely strong preferences not to live near African-Americans, a preference likely to be reflected in home values in the neighborhood. Indeed, the strong preference of whites not to live near African-Americans has by no means disappeared (Charles 2003; Bobo et al. 2012). The issue is whether the institutionalization of these preferences into a mortgage system by HOLC/FHA causally worsened racial segregation or lowered homeownership rates among African-Americans. Recent evidence suggests a positive answer to both questions.

In a working paper, economists from the Federal Reserve Bank of Chicago examine the causal impact of the Home Owners Loan Corporation redline maps by employing a number of identification strategies: they study changes over time in outcomes between neighbors living close to one another but at either side of an HOLC boundary; they examine separately those HOLC borders least likely to have been drawn endogenously; and they exploit the fact that the HOLC limited its maps to cities with a population of at least 40,000 by comparing findings in cities just below and just above that threshold (Aaronson, Hartley, and Mazumder 2019). The results are consistent with the HOLC boundaries having a causal impact on both racial segregation and lower outcomes for predominantly black neighborhoods. As the authors write, "areas graded 'D' become more heavily African-American than nearby C-rated areas over the 20th century, [a] . . . segregation gap [that] rises steadily from 1930 until about 1970 or 1980 before declining thereafter" (p. 3). They find a similar pattern when comparing C and B neighborhoods, even though "there were virtually no black residents in either C or B neighborhoods prior to the maps" (p. 3). Furthermore, the authors find "an economically important negative effect on homeownership, house values, rents, and vacancy rates with analogous time patterns to share African-American, suggesting economically significant housing disinvestment in the wake of restricted credit access" (pp. 2–3).

Though redlining eventually became illegal, the long-term consequences of these and other obstacles to homeownership for the black-white wealth gap, and for socioeconomic inequality more generally, surely lasted much longer, as the work of Aaronson, Hartley, and Mazumder (2019) makes clear. Because wealth accumulation among average Americans during the second half of the twentieth century resulted in a large measure from real estate appreciation, groups with greater access to housing credit in earlier periods accumulated more wealth. Furthermore, those homes could be used as collateral for educational loans or else passed onto children, further contributing to the racial wealth gap. As a result, over that period, white homeowners and their children experienced the substantial head start of cheap, government-funded loans that were effectively either unavailable to African-Americans or available only under less favorable terms.

Historic Discrimination Has Contemporary Consequences (via Laws)

The federal laws that created the Home Owners Loan Corporation and the Federal Housing Administration were important to both current and past racial disparities in wealth, because of the institutional practices that these particular organizations developed, enforced, and disseminated. But laws can also have a direct institutional impact independent of the creation of any organization, an impact that, again, can last multiple generations. The contemporary consequences of past discrimination can also be institutional in nature in this different way, directly through the law.

A notable case occurs when a law had explicit racial intent originally—that is, when it emerged from taste- or prejudice-based discrimination—but remains on the books as a race-neutral law that largely affects the same population. Though overt discrimination has been outlawed in many contexts over the years, including by several constitutional amendments and by the Civil Rights Act of 1964, laws that were explicitly and openly animated by racial prejudice in the past remain on the books. One of the most important cases relates to voting rights.

Many state laws currently disenfranchise imprisoned felons or people who have ever been convicted of a felony. Felon disenfranchisement laws today disproportionally affect African-Americans. These laws do not mention race and, in fact, are consistent with the US Constitution—abridgment of the right to vote as a result of "participation in rebellion, or other crime" was part of the Fourteenth Amendment. However, these laws rose dramatically in number and scope after the Civil War, following Reconstruction and the ratification of the Fifteenth Amendment, which gave African-Americans the right to vote (Holloway 2009, 2013).

At the time, many white politicians openly debated ways of countering what they considered the threat of the rising political power of African-Americans. Strategies included poll taxes, intimidation, illiteracy tests, and many others. In state constitutional conventions, for example, a frequent topic of debate was how to restrict the black vote legally. As a Mississippi political leader would recall in 1905 Congressional testimony: "When I was a rather young man Mississippi was trying to get up some sort of constitution that would get rid of the ignorant negro vote. Of course they had to get up something entirely fair on all hands . . ." (as quoted in Holloway 2013, 84). A 1894 editorial in a South Carolina newspaper made the point explicitly in its support for a state constitutional convention: "Fortunately, the opportunity is offered the white people of the State in the coming election to obviate all future danger and fortify the Anglo-Saxon civilization against every assault from within and without, and that is the calling of a constitutional convention to deal with the all important question of suffrage" (as quoted in Behrens, Uggen, and Manza 2003, 570).

Among the strategies for restricting the black vote, focusing on the voting rights of those convicted of crimes would have strong odds of surviving legal challenges because of the mention of "crime" in the Fourteenth Amendment. Changing the classification of crimes with an eye to the racial composition of common perpetrators was a common strategy.[3] For example, the 1901 Alabama constitutional convention, as Behrens, Uggen, and Manza (2003, 569) explain, "altered that state's felon disenfranchisement law to include all crimes of 'moral turpitude,' applying to misdemeanors and even to acts not punishable by law." At the time, though the laws themselves could not mention race, those debating them certainly could, and again they did so, often with remarkable lack of ambiguity. In his opening, presidential address to that Alabama convention, John B. Knox made the objectives clear: "[In

[3] Readers may note contemporary parallels in the disparities in punishments for crimes involving crack cocaine versus powdered cocaine.

1861], as now, the negro was the prominent factor in the issue. . . . And what is it that we want to do? Why it is within the limits imposed by the Federal Constitution, to establish white supremacy in this State. . . . The justification for whatever manipulation of the ballot that has occurred in this State has been the menace of negro domination . . ." (as quoted in Behrens, Uggen, and Manza 2003, 571).

To be clear, the full set of historical causes behind felon disenfranchisement is a large and complex topic, and we do not pretend to do it justice here (for a historical account, see Holloway 2013). However, as the first-hand accounts from Mississippi, South Carolina, and Alabama make clear, even if other factors were at play, the intent of many such laws was at least in large part expressly racial—to counter the threat that many whites saw from black political empowerment.

These first-hand accounts from a handful of states are also consistent with the national patterns over time in the institution of laws restricting the vote. Behrens, Uggen, and Manza (2003) examined the occurrence and timing of all felon disenfranchisement laws from post-Civil War Reconstruction to the present. Racially motivated legislators will be more likely to pass felon disenfranchisement laws if such laws are disproportionately likely to affect blacks. In a careful examination of the trends, Behrens, Uggen, and Manza (2003) found that, net of other factors, the probability of a first disenfranchisement law increased as the black prison population (their proxy for the black felon population) increased: "Each 1% increase in the percentage of prisoners who are nonwhite increases the odds by about 10% that a state will pass its first felon disenfranchisement law" (p. 586). They observed a similar pattern in the disenfranchisement of *ex*-felons; in their preferred model, net of other factors, "a 10% increase in a state's nonwhite prison population raises the odds of passing an ex-felon disenfranchisement law by almost 50% . . ." (p. 588).

The case of felon disenfranchisement illustrates that institutional discrimination can be perpetrated not only by organizations such as employers or real estate agencies, but also by the law. In many states, such laws remain in place. Behrens, Uggen, and Manza (2003) note that states slowly liberalized such laws beginning in the second half of the twentieth century by, for example, reinstating voting rights one or two years after an offender has served the term or eliminating felon or ex-felon voting prohibitions altogether. Still, as late as 2016, 48 states disenfranchised felons currently in prison, and many of these states also disenfranchised inmates, parolees, or probationers (Uggen, Larson, and Shannon 2016).

Two final points are worth noting. First, the pattern of voting-rights reinstatement is also consistent with the racial threat hypothesis: net of other factors, the proportion black in the prison population varies negatively with reinstatement of ex-felon voting rights (Behrens, Uggen, and Manza 2003). (However, the proportion black in the state is positively associated with reinstatement of ex-felon voting rights.) Second, such voting rights can affect not only African-Americans but also the nation as a whole. For example, in the year 2000, given Florida's disenfranchisement laws and high numbers of felons—more than 800,000 disenfranchised felons and ex-felons in the state at the time—in the absence of those laws it is possible that

Al Gore would have carried the state and, thus, won the presidency (Uggen and Manza 2002; for a contrasting view, see Burch 2012).

Ostensibly Minor Forms of Discrimination Can Have Important Consequences

Clearly, sociologists tend to approach the study of discrimination expansively. This expansiveness affects both the domains in which sociologists study discrimination and the way they think about its consequences. The domains have varied widely. Although sociologists have studied discrimination in job, housing, consumer, and credit markets, they have also examined it in contexts where the economic consequences are less clear or direct, such as dating and marriage markets, or in contexts of ordinary social interaction, such as entertainment venues, social clubs, and schools. These studies have used audit, survey, and ethnographic methods to examine questions as varied as whether black patrons are more likely to be denied entry to nightclubs, which behavior by black students is likely to be categorized as problematic by teachers, how black customers are treated in retail shops, and how black women fare in comparison to others in online dating sites (for example, Feagin and Sikes 1994; Lin and Lundquist 2013; May and Goldsmith 2018).

It is not difficult to see that some of these forms of discrimination can have consequences for economic inequality. For example, both the act of getting married and the socioeconomic status of one's spouse will affect income and wealth accumulation. Moreover, online dating sites today are responsible for an increasing proportion of marriages.

But not all forms of discrimination can be easily and directly traced to an important economic outcome. Sociologists continue to study such questions in part because discrimination is consequential not merely episodically but also cumulatively, not just at critical junctures but also over the slow, lifelong buildup of its everyday stings (National Research Council 2004). Researchers have argued that everyday discrimination can happen so repeatedly that such events eventually come to have a cumulative effect. Being repeatedly followed by a security guard at a store, repeatedly seated in an undesirable part of a restaurant, repeatedly confronted with racial slights at work, and other forms of discrimination that may seem trivial when considered individually constitute what psychiatrists have called "micro-aggressions" (Pierce 1970, 263; Sue et al. 2007; see also Feagin and Sikes 1994; Lacy 2007). Repeatedly experiencing these slights, insults, and individually minor conflicts is expected to eventually affect mental health and physical well-being.

Although episodic discrimination in contexts such as looking for a job is probably more important for economic outcomes, cumulative discrimination in everyday contexts might be more important for health outcomes. An associational public health study based on the 1995 Detroit Area Survey suggests this possibility. Williams et al. (1997) examined four standard measures of health: self-reported health, overall well-being, psychological distress, and number of days in the previous month incapacitated for health reasons. They asked whether these measures of health were associated with two measures of race-related stress: "discrimination," which referred to major experiences of unfair treatment during hiring, promotion, or interactions with the police; and "everyday discrimination," which, via a nine-item measure, captured "chronic, routine, and relatively minor experiences of unfair treatment." The everyday discrimination measure captured the frequency of experiences, such as "receiving poorer service than others in restaurants or stores" and "people acting as if you are not smart" (p. 340). After adjusting for demographic, socioeconomic, and health-related factors, the first measure—of major experiences of discrimination—was not significantly related to any of the four health outcomes, but the everyday discrimination measure was significantly associated with all of them. Moreover, it fully accounted for the difference between blacks and others in all of the measures except for psychological distress.

Of course, results of this kind are not dispositive. However, they make clear that everyday discrimination is distinct from what a job seeker might face before an employer; that minor but chronic experiences deserve attention in their own right; and that when studying the consequences of discrimination the focus should be expansive, including not only economic but also physical and mental health outcomes. In fact, multiple studies have uncovered associations between the experience of race discrimination and psychological distress, happiness and life satisfaction, self-esteem, and depression, among other outcomes (for reviews, see Williams, Neighbors, and Jackson 2003; Pascoe and Smart Richman 2009). Several studies have also found associations with physiological outcomes, such as high blood pressure (Krieger and Sidney 1996; Williams, Neighbors, and Jackson 2003, 200–201). There is clearly space for stronger work in this area, not only to determine better ways of modeling and testing for the effects of everyday discrimination, but also to uncover the mechanisms through which it matters.

Perceived Discrimination Is Important

The Williams et al. (1997) associational study of public health, like many in public health and sociology that assess the consequences of discrimination, focused on perceived discrimination. Such studies are concerned not with confirming whether discrimination has happened but with assessing how perceiving that it did matters. Economists are known for their healthy skepticism of studies that rely on what people say; actions, not words, many insist, are what matter. For many kinds of questions, we are inclined to agree. But the study of perceived discrimination is not a poor analytical substitute for that of perpetrated discrimination; it is the study of an entirely different question, the pursuit of which may reflect a difference between economics and other disciplines.

It would certainly be inappropriate to infer much about perpetrated discrimination from perceived discrimination. For example, researchers should not use changes in measures of perceived discrimination to assess whether discrimination has declined or risen. Sometimes, people perceive discrimination when it did not happen. Conversely, many victims of discrimination cannot have perceived that it took place. For example, a minority homeseeker cannot know either what units she would have been shown by the realtor or what terms she would have been offered by the banker had she been white.

Instead, the study of perceived discrimination is animated by a different set of concerns. For a potential victim's wages, odds of getting a job, mortgage rates, and other standard economic outcomes, perceiving whether discrimination happened is often immaterial—whether it actually took place is what matters. For a potential victim's mental health, depression, stress, and related health outcomes, perceiving that it happened is everything. Perceptions of discrimination can have an effect regardless of whether the perpetrator discriminated or instead seemed to discriminate but did not actually do so. If discrimination actually did happen but the potential victim did not perceive it, there may be little or no consequence for mental health and related outcomes.

To be clear, our point is not that economic outcomes and health outcomes inherently require one or the other perspective. For some economic outcomes, perception alone can matter, too. If people perceive discrimination and therefore withdraw from the job market, perception itself matters independent of actual discrimination. Conversely, if doctors treat black patients less attentively than white ones, then the health of black patients may suffer, regardless of whether the patients perceive the discriminatory difference in treatment. Our point is that studying perception is important entirely independently of studying actually perpetrated discrimination.

Note the implication of our discussion: whether people are right that they experienced discrimination will not matter at the two extremes—when actual discrimination is the sole concern, and when perception alone is. However, whether people are accurate in their perception of discrimination can also matter a great deal for two different kinds of questions. First, it can be important to understand why people perceive discrimination in spite of substantial evidence to the contrary. For example, some proportion of whites believe whites to be the racial group most discriminated against in the United States, a belief often accompanied by high levels of resentment. Investigating beliefs of this kind can be important in understanding people's political and social behavior. Two, it can be important to understand, particularly for health-related outcomes, how people respond to ambiguity—in this context, when potential victims are uncertain whether they have actually experienced discrimination. There are times when people do not know, but still wonder, whether they have been treated poorly because of the color of their skin. This possibility, if it happens repeatedly or in a consequential context, can cause a kind of rumination whose emotional strain can be taxing (Feagin and Sikes 1994; Lacy 2007). Efforts to measure the incidence and consequences of these everyday forms of discrimination more systematically would represent a particularly useful path forward.

Conclusion

We have argued that in addition to taste and statistical discrimination (and the possibility of implicit discrimination), economists should examine—or continue to examine—institutional discrimination. We have shown that institutional discrimination can take at least two forms, organizational and legal, and that in both forms the decisions of a contemporary actor to discriminate, whether out of animus or statistical averaging, can be immaterial. We have suggested that institutional discrimination is a vehicle through which past discrimination (intentional or not) has contemporary consequences. We have proposed that the minor forms of everyday discrimination people may experience deserve attention, because discrimination can matter cumulatively, not just episodically. And we have posited that the perception of discrimination is an important, independent topic deserving serious attention.

We stress that ours is not a comprehensive review of sociological perspectives on discrimination. We have ignored questions that researchers in that field may consider important. And not all sociologists would discuss the topics we have covered here as we have. However, we believe the topics discussed here deserve attention, and we are convinced that greater interaction between the two disciplines on these questions will benefit both.

■ *We thank Allison Daminger for excellent research assistance; Tara Garcia Mathewson and Mike Luca for conversations that have benefited this manuscript; Lincoln Quillian and András Tilcsik for valuable comments on an earlier draft; and Harvard University and the Harvard Project on Race, Class, and Cumulative Adversity at the Hutchins Center for generous support.*

References

Aaronson, Daniel, Daniel Hartley, and Bhashkar Mazumder. 2019. "The Effects of the 1930s HOLC 'Redlining' Maps." Federal Bank of Chicago Working Paper 2017-12.

Arrow, Kenneth J. 1972a. "Models of Job Discrimination." In *Racial Discrimination in Economic Life,* edited by A. Pascal, 83–102. Lexington: Lexington Heath.

Arrow, Kenneth J. 1972b. "Some Mathematical Models of Race Discrimination in the Labor Market." In *Racial Discrimination in Economic Life,* edited by A. Pascal, 187–203. Lexington: Lexington Heath.

Arrow, Kenneth J. 1998. "What Has Economics to Say about Racial Discrimination?" *Journal of Economic Perspectives* 12 (2): 91–100.

Becker, Gary S. 1971. *The Economics of Discrimination.* Chicago: University of Chicago Press.

Behrens, Angela, Christopher Uggen, and Jeff Manza. 2003. "Ballot Manipulation and the 'Menace of

Negro Domination': Racial Threat and Felon Disenfranchisement in the United States, 1850–2002." *American Journal of Sociology* 109 (3): 559–605.

Bertrand, Marianne, Dolly Chugh, and Sendhil Mullainathan. 2005. "Implicit Discrimination." *American Economic Review* 95 (2): 94–98.

Bielby, William T. 2000. "Minimizing Workplace Gender and Racial Bias." *Contemporary Sociology* 29 (1): 120–29.

Bobo, Lawrence D., Camille Z. Charles, Maria Krysan, and Alicia D. Simmons. 2012. "The Real Record on Racial Attitudes." In *Social Trends in American Life: Finds from the General Social Survey since 1972*, edited by Peter V. Marsden, 38–83. Princeton: Princeton University Press.

Burch, Traci. 2012. "Did Disfranchisement Laws Help Elect President Bush? New Evidence on the Turnout Rates and Candidate Preferences of Florida's Ex-Felons." *Political Behavior* 34 (1): 1–26.

Cancio, A. Silvia, David T. Evans, and David J. Maume, Jr. 1996. "Reconsidering the Declining Significance of Race: Racial Differences in Early Career Wages." *American Sociological Review* 61 (4): 541–56.

Charles, Camille Zubrinsky. 2003. "The Dynamics of Racial Residential Segregation." *Annual Review of Sociology* 29: 167–207.

Crossney, Kristen B., and David W. Bartelt. 2005. "The Legacy of the Home Owners' Loan Corporation." *Housing Policy Debate* 16 (3–4): 547–74.

Dobbin, Frank. 2009. *Inventing Equal Opportunity*. Princeton: Princeton University Press.

Dobbin, Frank, Daniel Schrage, and Alexandra Kalev. 2015. "Resisting against the Iron Cage: The Varied Effects of Bureaucratic Personnel Reforms on Diversity." *American Sociological Review* 80 (5): 1014–44.

Elliott, James R., and Ryan A. Smith. 2004. "Race, Gender, and Workplace Power." *American Sociological Review* 69 (3): 365–86.

Farkas, George, and Keven Vicknair. 1996. "Appropriate Tests of Racial Wage Discrimination Require Controls for Cognitive Skill: Comment on Cancio, Evans, and Maume." *American Sociological Review* 61 (4): 557–60.

Feagin, Joe R., and Douglas Lee Eckberg. 1980. "Discrimination: Motivation, Action, Effects, and Context." *Annual Review of Sociology* 6 (1): 1–20.

Feagin, Joe R., and Melvin P. Sikes. 1994. *Living with Racism: The Black Middle-Class Experience*. Boston: Beacon Press.

Fernandez, Roberto M., and Isabel Fernandez-Mateo. 2006. "Networks, Race, and Hiring." *American Sociological Review* 71 (1): 42–71.

Fernandez-Mateo, Isabel. 2009. "Cumulative Gender Disadvantage in Contract Employment." *American Journal of Sociology* 114 (4): 871–923.

Gaddis, S. Michael. 2015. "Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market." *Social Forces* 93 (4): 1451–79.

Guryan, Jonathan, and Kerwin Kofi Charles. 2013. "Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots." *Economic Journal* 123 (572): F417–32.

Hillier, Amy E. 2003. "Redlining and the Home Owners' Loan Corporation." *Journal of Urban History* 29 (4): 394–420.

Hillier, Amy E. 2005. "Residential Security Maps and Neighborhood Appraisals: The Home Owners' Loan Corporation and the Case of Philadelphia." *Social Science History* 29 (2): 207–33.

Holloway, Pippa. 2009. "'A Chicken-Stealer Shall Lose His Vote': Disfranchisement for Larceny in the South, 1874–1890." *Journal of Southern History* 75 (4): 931–62.

Holloway, Pippa. 2013. *Living in Infamy: Felon Disfranchisement and the History of American Citizenship*. New York: Oxford University Press.

Hultin, Mia, and Ryszard Szulkin. 1999. "Wages and Unequal Access to Organizational Power: An Empirical Test of Gender Discrimination." *Administrative Science Quarterly* 44 (3): 453–72.

Jackson, Kenneth T. 1980. "Race, Ethnicity, and Real Estate Appraisal: The Home Owners Loan Corporation and the Federal Housing Administration." *Journal of Urban History* 6 (4): 419–52.

Jackson, Kenneth T. 1985. *Crabgrass Frontier: The Suburbanization of the United States*. New York: Oxford University Press.

Kalev, Alexandra. 2014. "How You Downsize Is Who You Downsize: Biased Formalization, Accountability, and Managerial Diversity." *American Sociological Review* 79 (1): 109–35.

Kang, Sonia K., Katherine A. DeCelles, András Tilcsik, and Sora Jun. 2016. "Whitened Résumés: Race and Self-Presentation in the Labor Market." *Administrative Science Quarterly* 61 (3): 469–502.

Krieger, Nancy, and Stephen Sidney. 1996. "Racial Discrimination and Blood Pressure: The CARDIA

Study of Young Black and White Adults." *American Journal of Public Health* 86 (10): 1370–78.

Lacy, Karyn R. 2007. *Blue-Chip Black: Race, Class, and Status in the New Black Middle Class.* Berkeley: University of California Press.

Light, Ryan, Vincent J. Roscigno, and Alexandra Kalev. 2011. "Racial Discrimination, Interpretation, and Legitimation at Work." *ANNALS of the American Academy of Political and Social Science* 634 (1): 39-59.

Lin, Ken-Hou, and Jennifer Lundquist. 2013. "Mate Selection in Cyberspace: The Intersection of Race, Gender, and Education." *American Journal of Sociology* 119 (1): 183–215.

Lucas, Samuel Roundfield. 2008. *Theorizing Discrimination in an Era of Contested Prejudice: Discrimination in the United States.* Philadelphia: Temple University Press.

Massey, Douglas S., and Nancy A. Denton. 1993. *American Apartheid: Segregation and the Making of the Underclass.* Cambridge: Harvard University Press.

May, Reuben A. Buford, and Pat Rubio Goldsmith. 2018. "Dress Codes and Racial Discrimination in Urban Nightclubs." *Sociology of Race and Ethnicity* 4 (4): 555–66.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27: 415–44.

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* 2nd ed. Cambridge: Cambridge University Press.

Mouw, Ted. 2002. "Are Black Workers Missing the Connection? The Effect of Spatial Distance and Employee Referrals on Interfirm Racial Segregation." *Demography* 39 (3): 507–28.

National Research Council. 2004. *Measuring Racial Discrimination.* Washington, DC: National Academies Press.

Norris, David, and Mikyung Baek. 2016. *Full Report: H.E.A.T.* Columbus: Kirwan Institute for the Study of Race and Ethnicity.

Oliver, Melvin L., and Thomas M. Shapiro. 2006. *Black Wealth/White Wealth: A New Perspective on Racial Inequality.* New York: Routledge.

Pager, Devah. 2007. "The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future." *ANNALS of the American Academy of Political and Social Science* 609: 104–33.

Pager, Devah, and Diana Karafin. 2009. "Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision Making." *ANNALS of the American Academy of Political and Social Science* 621: 70–93.

Pager, Devah, and Hana Shepherd. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Review of Sociology* 34 (1): 181–209.

Pascoe, Elizabeth A., and Laura Smart Richman. 2009. "Perceived Discrimination and Health: A Meta-Analytic Review." *Psychological Bulletin* 135 (4): 531–54.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62 (4): 659–61.

Pierce, Chester. 1970. "Offensive Mechanisms." In *The Black Seventies,* edited by Floyd Barbour. Boston: Porter Sargent.

Powell, Walter W., and Paul J. DiMaggio. 1991. *The New Institutionalism in Organizational Analysis.* Chicago: University of Chicago Press.

Quillian, Lincoln, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen. 2017. "Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time." *Proceedings of the National Academy of Sciences* 114 (41): 10870–75.

Reskin, Barbara F. 2000. "The Proximate Causes of Employment Discrimination." *Contemporary Sociology* 29 (2): 319–28.

Reskin, Barbara. 2012. "The Race Discrimination System." *Annual Review of Sociology* 38: 17–35.

Rubineau, Brian, and Roberto M. Fernandez. 2013. "Missing Links: Referrer Behavior and Job Segregation." *Management Science* 59 (11): 2470–89.

Scott, W. Richard. 2013. *Institutions and Organizations: Ideas, Interests, and Identities.* 4th ed. Thousand Oaks: Sage Publications.

Small, Mario Luis. 2009. *Unanticipated Gains: Origins of Network Inequality in Everyday Life.* Oxford: Oxford University Press.

Small, Mario L. 2011. "How to Conduct a Mixed Method Study: Recent Trends in a Rapidly Growing Literature." *Annual Review of Sociology* 37: 57–86.

Small, Mario L., David J. Harding, and Michèle Lamont. 2010. "Reconsidering Culture and Poverty." *ANNALS of the American Academy of Political and Social Science* 629 (1): 6–27.

Sue, Derald Wing, Christina M. Capodilupo, Gina C. Torino, Jennifer M. Bucceri, Aisha M.B. Holder, Kevin L. Nadal, and Marta Esquilin. 2007. "Racial Microaggressions in Everyday Life: Implications for Clinical Practice." *American Psychologist* 62 (4): 271–86.

Sugrue, Thomas J. 1996. *The Origins of the Urban Crisis: Race and Inequality in Postwar Detroit.* Princeton: Princeton University Press.

Turco, Catherine J. 2010. "Cultural Foundations of Tokenism: Evidence from the Leveraged Buyout Industry." *American Sociological Review* 75 (6): 894–913.

Uggen, Christopher, Ryan Larson, and Sarah Shannon. 2016. *6 Million Lost Voters: State-Level Estimates of Felony Disenfranchisement, 2016.* Washington, DC: The Sentencing Project.

Uggen, Christopher, and Jeff Manza. 2002. "Democratic Contraction? Political Consequences of Felon Disenfranchisement in the United States." *American Sociological Review* 67 (6): 777–803.

Waldinger, Roger, and Michael L. Lichter. 2003. *How the Other Half Works: Immigration and the Social Organization of Labor.* Berkeley: University of California Press.

Williams, David R., Harold W. Neighbors, and James S. Jackson. 2003. "Racial/Ethnic Discrimination and Health: Findings from Community Studies." *American Journal of Public Health* 93 (2): 200–208.

Williams, David R., Yan Yu, James S. Jackson, and Norman B. Anderson. 1997. "Racial Differences in Physical and Mental Health: Socio-economic Status, Stress and Discrimination." *Journal of Health Psychology* 2 (3): 335–51.

Wilson, William Julius. 1978. *The Declining Significance of Race: Blacks and Changing American Institutions.* Chicago: University of Chicago Press.

Wilson, William Julius. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy.* Chicago: University of Chicago Press.

Wilson, William Julius. 1996. *When Work Disappears: The World of the New Urban Poor.* New York: Knopf.

Journal of Economic Perspectives—Volume 34, Number 2—Spring 2020—Pages 68–89

Race Discrimination: An Economic Perspective

Kevin Lang and Ariella Kahn-Lang Spitzer

In this article, we discuss the theory and evidence on discrimination in two key domains—the labor market and the criminal justice system—from an economic perspective. We define discrimination as treating someone differently based on characteristics such as gender, race, or religion. Prejudice may lead to discrimination, but only if you act on it. Moreover, discrimination by individuals does not necessarily lead to discrimination at a market or societal level. We focus primarily on discrimination against blacks, although many of the concepts discussed apply much more broadly.

While documenting racial disparities is relatively easy, identifying discrimination as the cause is more challenging. Discrimination may create racial disparities in outcomes, but so can differences in preferences or true underlying differences in innate characteristics. Thus, we begin with two sections discussing approaches to identifying discrimination in the labor market and criminal justice system. One theme of this discussion is that discrimination can happen in different areas. For example, unequal labor market outcomes could result from discrimination by employers, or discrimination by potential coworkers, or the discriminatory attitudes of customers, or some combination of these. Unequal outcomes in the criminal justice system could result from discrimination in police actions or court decisions. Of course, focusing on the areas in which discrimination occurs can help us better focus antidiscrimination policy.

■ *Kevin Lang is Professor of Economics, Boston University, Boston, Massachusetts. Ariella Kahn-Lang Spitzer is a Human Services Researcher, Mathematica Policy Research, Cambridge, Massachusetts. Their email addresses are lang@bu.edu and AKahn-Lang@mathematica-mpr.com.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.2.68.

Even when disparities can be attributed to discrimination, the causes of discriminatory behavior may differ. Economists distinguish between two main models: "taste-based" and "statistical" discrimination. For each of these models, we take a deeper look at both the theory and the evidence.

Taste-based discrimination reflects prejudice or preferences. Thus, an employer who hires men rather than women because of a personal preference for working with men is discriminating based on tastes. Taste-based discrimination can also reflect *invalid* statistical inference. Thus, someone who, contrary to a large body of evidence, believes immigrants are more likely to commit violent crimes is discriminating based on prejudice. Note that distinguishing between valid and invalid statistical discrimination is not always straightforward. For example, invalid statistical inference may reflect valid statistical assessment of a nonrepresentative sample, as when an employer ascribes differences in an earlier job applicants pool to a current one even though the applicant population has changed.

The canonical Becker ([1957] 1971) model of employer discrimination suggests that market forces push back against taste-based discrimination, because prejudiced employers will pay more for (or hire less qualified) preferred race workers, thereby decreasing profits. Consequently, in a basic model, competition from nondiscriminating firms drives discriminators from the market until wage differentials between equally productive workers are eliminated. Becker's theory suggests that taste-based discrimination is most likely where 1) the race of the worker is salient and the customer market is not easily segmented or 2) the forces of competition are weak or absent. The second condition justifies focusing on areas such as law enforcement and criminal justice that are largely immune from competitive forces. Given this theoretical perspective, it is not surprising that there is considerable (although not universal) evidence of race discrimination in the US justice system at virtually every stage. It is, perhaps, more surprising that there is also evidence of taste-based labor market discrimination.

Statistical discrimination, first developed in the pioneering work of Phelps (1972) and Arrow (1972a, b), is discrimination based on *valid* statistical inference. For example, doctors typically discriminate between men and women regarding who should receive breast cancer screening. Although some men do get breast cancer, it is much more common among women; therefore, doctors typically recommend screening for women but not men. When a characteristic like race is correlated with unobserved or imperfectly observed productivity, criminal conduct, or some other factor, people may use that characteristic to update their prior estimates. Firms engaging in statistical discrimination maximize profits; actors in the law enforcement and legal systems may be acting as rational Bayesians. But even when based on valid statistical inference, this form of discrimination can be harmful to individuals and socially undesirable.

Because informational imperfections drive statistical discrimination, firms may seek additional information. We discuss and critically assess the effect of increasing information on discrimination. The role of information is a particularly salient issue in criminal justice. Risk prediction algorithms seemingly remove implicit and explicit bias from sentencing and bail decisions. However, discrimination in criminal justice treatment can lead to discriminatory algorithms in a manner analogous to the role of disparities in promoting statistical discrimination.

In the conclusion, we point out that although economic studies often focus on discrimination in specific domains like the labor market or the criminal justice system, real-world discrimination arises in a system of self-reinforcing linkages between different domains of discrimination. For example, discrimination leads to social and residential distance, which reinforces between-group differences. Such differences, in turn, favor additional discrimination and social distance. This suggests that policies to address discrimination might usefully seek key points of leverage that could propagate through a range of outcomes.

Discrimination in the Labor Market

Racial disparities in the labor market are readily apparent. As one example, black men in 2010, relative to white men, were 28 percent less likely to be employed and earned 31 percent less annually conditional on employment (Kahn-Lang 2018). Relative to white women, black women also earn less, although the differential is only about half that for males (Daly, Hobijn, and Pedtke 2020). This is in part because it is obscured by a strong positive relation between skill and employment among black but not white women (Neal 2004). These disparities do not prove that labor market discrimination exists, but they surely suggest that the question is worth exploring.

Traditionally, much of the evidence for labor market discrimination came from ordinary least squares regressions with wages as the dependent variable and a set of control variables like age and education, along with a dummy variable for race, as the explanatory variables. The working assumption was that if the coefficient on the dummy variable for race was significant, this was evidence of discrimination. However, because the set of observable control variables for differences between blacks and whites was inevitably incomplete, at best the coefficient on race represented an unexplained differential that might reflect discrimination. In addition, some control variables might themselves reflect past discrimination. For example, using parental income as a control variable reduces the racial gap in earnings. However, lower levels of parental income for blacks relative to whites might reflect labor market discrimination experienced by their parents.

As an example of the controversies that arise around these kinds of results, Neal and Johnson (1996) showed that controlling for performance on the Armed Forces Qualifying Test (AFQT), a measure of cognitive skill, eliminated roughly three-quarters of the black-white wage differential among men. However, Lang and Manove (2011) showed that, conditional on their AFQT score, blacks get more education than whites do. Adding education to the controls raises the estimated black-white wage differential by about six percentage points. Moreover, this gap persists with an added “kitchen sink” of control variables, although it remains possible that including some other variable would eliminate the gap.

One useful perspective from Becker ([1957] 1971) is that there are multiple possible sources of taste-based discrimination in hiring—the employer, coworkers, and customers—and their effects would not be the same. In addition, focusing on specific decision-makers allows for research methods that provide a cleaner empirical test for the existence of discrimination.

Employer Discrimination

Researchers have proposed a number of strategies for identifying discrimination by employers. Goldin and Rouse (2000) take advantage of a natural experiment hiring in which symphony orchestras switched to blind auditions. They find that females are substantially more likely to be hired when auditions are blind than when employers observe gender during auditions. Unfortunately, such natural experiments are rare. Consequently, most researchers have relied on experimental evidence to identify discrimination.

"Audit studies," in which matched pairs of black and white actors posing as workers and using similar fictitious resumes applied for jobs, have been used to study employer-based discrimination. Bendick (2007) reviewed ten such studies; all showed disparate treatment favoring whites, although some disparities were statistically insignificant. However, there are concerns that these studies may pick up something other than discrimination. Audit studies attempt to match the black and white applicants as closely as possible but cannot match them perfectly. Therefore, researchers might accidentally or unconsciously choose white applicants whose appearance or presentation make them more attractive to employers for reasons unrelated to race. Further, actors who know they are in a study may subconsciously act differently based on their role.

To avoid this problem, researchers turned to "correspondence studies" in which information on the application—most commonly name—signals race. Because the resumes are fictional, researchers can ensure that, except for the information signaling race, the content of resumes is uncorrelated with race. In a well-known study, Bertrand and Mullainathan (2004) found that 9.7 percent of resumes with a white-sounding name elicited a callback, relative to 6.5 percent of those with black-sounding names. However, name may signal more than race; for example, it might also signal social class. Jacquemet and Yannelis (2012) find support for this concern, showing considerable variation in callback rates across names within race. However, Fryer and Levitt (2004) find that blacks with and without black-sounding names have similar outcomes once they control for zip code at birth, suggesting that either discrimination is based on race, not names, or discrimination at the resume-screening stage may not translate to discriminatory outcomes.

Some recent papers use within-establishment variation to make a more compelling case for discrimination. Giuliano, Levine, and Leonard (2009) studied a large US retailer with many outlets. Because hiring managers in these outlets change frequently, the authors could compare hiring by black and white managers in the same outlet. Relative to black managers, white managers hired more white workers and fewer black workers, especially in the South. This pattern could reflect a number of decision processes: 1) discrimination by white and/or black managers, 2) synergies between same-race managers and workers, 3) different hiring networks, or 4) workers preferring managers of their own race. The evidence for each of the last three is weak. For example, the relative performance of black workers is higher under black managers, but the estimate falls well short of statistical significance at conventional levels. Managers are more likely to hire workers who live near them, but this accounts for little of the own-race effect. White workers are somewhat more likely to quit when they get a new black manager, but the estimated effect is only marginally statistically significant. Discrimination favoring own-race employees seems to us to be the dominant explanation in this study, although admittedly this is the residual explanation.

Coworker Discrimination

Perhaps employers do not themselves have a taste for discrimination but are pressured to act as if they do because of prejudice from their employees. As one example, the Giuliano, Levine, and Leonard (2009) study above also provides evidence that workers are less likely to quit when more coworkers have the same race/ethnicity. These effects are large and highly significant for whites and Asians, smaller and marginally significant for blacks, and small and statistically insignificant for Hispanics.

In contrast, Bygren (2010) finds, in a matched sample of Swedish firms and workers, that workers are less likely to leave a given establishment when there are more workers of the *opposite* sex, suggesting at least that Swedish workers do not systematically prefer to work with their own sex.

In an experimental study, Hedegaard and Tyran (2018) hired secondary school students for real, albeit short-term, jobs preparing letters for mailing. Initially, workers worked alone and were paid piece rate. They were then told that they would work in pairs that shared compensation. Some workers could choose between a worker with a Danish-sounding name and one with a Muslim-sounding name. They were also told how many letters each worker had prepared the previous week. Assuming workers believe that last week's output predicts output when paired with another worker, the authors calculated the cost of choosing the less productive worker. On average, workers were willing to pay 8 percent of their earnings for two days to work with someone with the same ethnicity.

Customer Discrimination

Still another possibility is that employers who are not themselves prejudiced discriminate in hiring because their customers are prejudiced. The importance of customer-based discrimination probably depends heavily on the product or service. Some relevant studies focus on intriguing groups from which it may be unwise to draw broad conclusions. In "fantasy" sports, Bryson and Chevalier (2015) find that, conditional on price and past performance, white and nonwhite players are equally likely to be selected when the season begins or traded during the season in the

(English) Fantasy Premier League. Similarly, Broyles and Keen (2010) find that trading card prices for players in the (American) National Basketball Association are unrelated to player race. These contexts, however, require no true interaction between "customer" and "player." In a much more intimate context, brothel owners in New York charged a premium for lighter-skin blacks and a larger premium for white sex workers (Mumford 1997, p. 105). From the other side of the market, Li, Lang, and Leong (2018) find evidence of discrimination against darker-skin customers by present-day Singapore street sex workers.

In practice, the intimacy of most interactions between customers and workers falls somewhere in the middle of that between fantasy sports teams and players and that between clients and sex workers, and the evidence on customer discrimination in more mainstream settings is limited. Customers give smaller tips to black taxi drivers than to white ones, but we cannot be sure that this disparity is unrelated to service quality (Ayres, Vars, and Zakariya 2005).

Leonard, Levine, and Giuliano (2010) used data from the retailer described above to assess the importance of customer discrimination in a more common setting. They found that in areas with a larger proportion of whites, having more black employees slightly reduces sales, but having more Hispanics slightly increases them; the results are small in either case and, given the large number of hypotheses tested, may be spurious. They do find benefits from having more Asian workers when the proportion of individuals nearby speaking only Asian-Pacific languages is high. Similarly, Combes et al. (2016) show that a higher proportion of French residents is associated with a larger increase in the disparity in employment between African and French workers in jobs with customer contact than in those without such contact. They argue that this is best explained by customer discrimination.

Perhaps the strongest evidence for customer discrimination stems from online transactions with individual sellers. Buyers were less likely to make an offer to purchase an iPod Nano (portable digital music player) offered by a black person and made lower offers if they did (Doleac and Stein 2013). Similarly, Arab sellers and buyers faced discrimination in an online market for used automobiles in Israel (Zussman 2013). The authors suggest that customers must trust that the product is legitimate, as advertised, and procured legally and that race or ethnicity affects perceived trustworthiness.

Linking Evidence on Discrimination to Broader Disparities

There is a missing link between the evidence that most clearly demonstrates the existence of labor market discrimination and the size of the racial disparities in labor markets. Many of the studies regarding discrimination have focused on a specific group and setting: a large US retailer, resumes submitted to a certain group of employers, online buyers, and so on. The narrow focus of these studies helps make their statistical identification persuasive but makes it harder to draw a direct connection to the aggregate racial disparities in labor markets. For example, even if some firms discriminate when screening resumes, it is unclear how this translates into employment and earnings disparities.

In addition, there are theoretical reasons to hesitate before jumping straight from evidence of discrimination by some to aggregate results. As the Becker ([1957] 1971) model points out, if only some firms discriminate by race, blacks can find equally desirable employment at other firms. Thus, workplaces could show a high degree of segregation by race without a resulting gap in wages.

Discrimination in the Criminal Justice System

Policing

There are clear racial disparities in the criminal justice system. Such discrepancies are particularly salient in policing. One estimate suggests that blacks and whites use marijuana at similar rates, but blacks are 3.7 times as likely to be arrested for its use (ACLU 2013). Similarly, black drivers are stopped more frequently than white drivers and are more likely to be subjected to search if stopped (Pierson et al. 2017).

What other factors might account, at least in part, for such discrepancies? Location is one possibility. Crime is more concentrated in black communities. This leads to increased policing in those locations, which may increase the likelihood that black drivers are stopped or arrested. However, racial disparities remain after accounting for location. Using data on state patrol stops in 20 states, Pierson et al. (2017) estimate that black drivers are 40 percent more likely to be stopped than white drivers, after controlling for age, gender, and location. As noted earlier, however, such disparities do not prove the existence of discrimination. The remaining disparities could reflect differences in driving behavior; black drivers may speed or break other traffic laws more frequently than whites do. Similarly, blacks may carry larger amounts of marijuana or use it in more public places. It is difficult to dismiss such possibilities, because we generally lack data on offenders who were not apprehended. Similarly, we observe these events in the data as documented by the police, who may also be biased by discrimination.

Again, the challenge is to find research techniques that provide evidence on whether disparities in policing are due to discrimination. Such studies have produced conflicting results. One approach, called the "outcomes model," argues that absent discrimination, black and white drivers on the margin of being stopped should be equally likely to be found at fault. If, conditional on a stop, blacks are less likely to be found at fault, this suggests discrimination. However, this insight only applies for the marginal person stopped, something which is typically unobservable; we cannot simply compare the average rates at which searches uncover contraband. Knowles, Persico, and Todd (2001) address this issue by modeling police searches during traffic stops as resulting from sequential decisions in which the driver first decides whether to carry contraband and the police decide whether to search based on the proportion of drivers with contraband. They show that in the equilibrium of this model, the average and marginal rates of contraband found during police searches will be equal. Using data on traffic stops in Maryland, they find similar rates of contraband on white and black drivers, conditional on search. They conclude that search differentials are consistent with no discrimination.

However, Engel and Tillyer (2008) argue that this method requires the strong assumption that drivers are rational actors with full information regarding the likelihood of being stopped and searched. Simoiu, Corbett-Davies, and Goel (2017) model the police decision to search as a function of a continuous signal, sent by drivers to police, on their likelihood of carrying contraband. They show that by imposing a strict functional form on the distribution of the signals, they can identify the police threshold for search. They find police have a lower signal threshold for search for black and Hispanic drivers relative to white drivers, suggesting the presence of discrimination.

Another approach argues that the “veil of darkness” at night makes it harder for police to discriminate based on race. In other words, if racial discrepancies reflect discrimination, they should be more prevalent during daylight hours. Grogger and Ridgeway (2006) find that at times of day that are dark only at certain times of the year, racial disparities in police stops are unrelated to whether it is dark. Horrace and Rohlin (2016) measure whether streets are well lit during nondaylight hours. After accounting for street lighting, they find light is associated with a 15 percent increase in the odds of a black driver being stopped relative to a white driver. Kalinowski, Ross, and Ross (2017) further argue that drivers may rationally respond to differences in police behavior in darkness. After accounting for this in a theoretical model, they find support for police discrimination.

Fryer (2019) finds that after controlling for key characteristics of police interactions, there are no racial discrepancies in officer-involved shootings. However, he finds that police are more likely to use force against blacks and Hispanics. In a working paper commenting on the results, Knox, Lowe, and Mummolo (2019) argue that Fryer’s estimates are likely understated because they do not account for bias in administrative police data. Police may be more likely to interact with or record interactions with blacks and, conditional on recording an interaction, may record more severe conditions. Assuming reasonable discrepancies in recording by race dramatically increases the estimated discriminatory component of force against blacks.

Courts

Court settings also show substantial disparities by race. Blacks are more likely to be assigned monetary bail instead of being released without bail, be assigned higher monetary bail conditional on getting monetary bail (Arnold, Dobbie, and Yang 2018), be convicted conditional on being charged (Anwar, Bayer, and Hjalmarsson 2012), and receive harsher sentences conditional on conviction (Mauer 2011). Once again, it is challenging to determine the extent to which this reflects discrimination rather than other factors. First, disparities may represent true differences in observable and unobservable case characteristics. Using a rich dataset with substantial case information, Rehavi and Starr (2014) show that controlling for measured case characteristics eliminates much, but not all, of the racial disparities in sentencing. In addition, black defendants, on average, have access to fewer resources than white defendants, which plausibly leads to inferior legal representation and ability to navigate the system.

Much of the research on identifying discrimination in court settings has relied on the outcomes model. Some judges are stricter, while others are more lenient. Consequently, some defendants receive bail only because they were randomly assigned to a lenient judge. Arnold, Dobbie, and Yang (2018) argue that using random judge assignment as an instrument for bail setting allows them to identify the marginal defendants—that is, those who would be granted monetary bail by more lenient judges but not released by others. They find less pretrial misconduct by marginally released black defendants than marginally released white defendants, which implies substantial discrimination in bail setting. In contrast, Anwar and Fang (2015) identify marginal parole applicants as applicants granted parole between their minimum and maximum sentences, arguing that because parole can be granted at any time in this range, prisoners will tend to be released at the point when marginal benefit equals marginal cost. They observe no racial disparity in recidivism among prisoners released by the parole board in this period and thus no evidence of discrimination.

There is limited clear evidence of discrimination in sentencing. Abrams, Bertrand, and Mullainathan (2012) show that despite random assignment of defendants to judges, the relative incarceration rates of black and white defendants vary among judges. Therefore, they argue that there is at least some discrimination in sentencing. They find no statistically significant variation in relative sentence lengths conditional on incarceration. Alesina and La Ferrare (2014) find that minority defendants' death sentences are overturned more frequently on appeal, suggesting discrimination in the lower courts. However, this conclusion requires that the superior courts have improved accuracy which, in turn, limits racial bias.

Taste-Based Discrimination

The Importance of Labor Market Frictions in the Theory

In the simplest version of Becker's ([1957] 1971) canonical model of employer discrimination, employers dislike hiring black workers and require a fixed level of compensation to hire a black worker rather than a white one. If the black-white wage gap exceeds this compensating differential, the employer hires only blacks; if not, the employer hires only whites. However, the prejudiced behavior by some firms means that less prejudiced firms hiring only black workers are more profitable, because they can hire productive workers at relatively low wages. The less prejudiced firms expand, while all-white firms contract. This increases the demand for black workers, and their relative wages rise. If there are sufficient unprejudiced employers, they will eventually drive the wage differential to zero. Prejudiced employers may survive and employ only white workers, but black and white workers will be equally well off.

The analysis is similar in the case of coworker bias. If white workers require a compensating differential for working alongside black workers, firms should hire either an all-white or an all-black workforce, but after the adjustments are done, there should be no wage gap. If customers dislike being served by black workers, a wage gap will persist only if there are too few unprejudiced consumers to be served by those black workers who are not employed in jobs where they are invisible to consumers.

Thus, one key takeaway from Becker's model is that the extent to which discrimination affects wages depends on the proportions of employers who are highly or mildly prejudiced and the flexibility of the market in allocating black workers to the least prejudiced firms. Lang and Lehmann (2012) argue that models of taste discrimination requiring a large number of highly discriminatory employers are inconsistent with survey evidence, which suggests that most Americans are not highly prejudiced.[1] We therefore limit our discussion to models of taste-based discrimination based on either a relatively small proportion of highly discriminatory employers or a large number of mildly prejudiced employers.

Taste-based discrimination models with labor market frictions generally assume that job applications have an opportunity cost. Therefore, workers do not apply for jobs they are unlikely to get or where they anticipate being unproductive. For example, Rosén (1997) assumes that each unemployed worker is matched with exactly one vacancy each period, but a vacancy may have multiple applicants. In this model, workers learn about their own match-specific productivity after being matched. If hired, a worker earns a fixed proportion of that match-specific productivity. Because the worker engages in sequential search with no recall, there is no on-the-job search, and there is a (very) small cost to bargaining, the worker applies only to jobs at which productivity exceeds a given reservation level. The firm sees the workers who choose to apply and selects one worker with whom to bargain. In the method of bargaining in the Rosén (1997) model ("Rubinstein bargaining"), the wage is a fixed proportion of match-specific productivity. Therefore, the firm wants to bargain with the worker with the highest match-specific productivity but does not observe this information, which is private to the worker.

Suppose that for some reason (there will be a reason in equilibrium, but we're not there yet) when both black and white workers apply, at least some firms choose to bargain first with whites. On average, black workers will have to search longer to find a job and will therefore set a lower reservation match-specific productivity. Consequently, firms that are otherwise indifferent between blacks and whites know blacks have a lower reservation productivity than whites. Therefore, firms would prefer to bargain with whites because they have higher expected productivity. If

[1] For example, fully 96 percent of Americans say they would be willing to vote for a black person for president. Doubtless, some survey respondents hide socially unacceptable feelings. However, in 2015 91 percent and in 2019, 95 percent of survey respondents said they would vote for a woman (McCarthy 2019). In contrast, using a list technique designed to eliminate social acceptability bias, Burden, Ono, and Yamada (2017) estimated that 13 percent would not vote for a woman. This suggests to us that while very low levels of expressed prejudice do underestimate true prejudice, they are not hiding very widespread strong prejudice.

no firm is prejudiced, there are only two stable equilibria: either all firms prefer to bargain with blacks or they all prefer to bargain with whites. If even a small group of firms is highly prejudiced against blacks in this setting, the equilibrium in which all firms discriminate against blacks seems more natural.

In the model of Lang, Manove, and Dickens (2005), firms announce wages simultaneously. Workers observe all the posted wages, and each applies to a single firm. If a firm receives at least one application, the firm hires one worker at the announced wage. Because the wage is fixed, if firms have a mild preference for white workers, they always choose a white applicant over a black one. Consequently, blacks strongly prefer not to apply where whites are likely to apply. In equilibrium, there are two wages, a high wage with a low vacancy rate attracting only whites and a low wage with a high vacancy rate attracting only blacks. With heterogeneous risk aversion, highly risk-averse whites may apply to the same jobs as less risk-averse blacks, in which case, such blacks will have relatively low rates of job finding. In this model, the discriminatory equilibrium is more plausible when mild prejudice is widespread.

In sum, when there are labor market frictions and wages are not set competitively, equally productive black workers may not be costlessly reallocated to alternative and equally paid jobs, while prejudiced firms may not have lower profits and therefore need not be driven out of business.

Evidence of Taste-Based Discrimination

Most people do not admit or may not recognize that they are discriminating, let alone attribute it to prejudice, making discrimination hard to identify and measure. This section describes some evidence on taste-based discrimination and the strategies that researchers have used to identify it.

Charles and Guryan (2008) use the simple version of the Becker ([1957] 1971) model to test for taste discrimination by employers. In this model, the racial prejudice of the marginal employer of black employees determines the racial wage gap. Because blacks represent a minority of workers, the racial prejudice of relatively unprejudiced employers—those hiring the marginal worker between the more prejudiced and the less prejudiced employers—should determine the wage gap. (Note that statistical discrimination models apply across all rational employers and thus do not make this prediction.) The authors use questions from the General Social Survey, such as whether the respondent opposes interracial marriage or would not vote for a black president, to create a "prejudice index." They then estimate the tenth, fiftieth, and ninetieth percentile of racial prejudice in each state. In one state, the median respondent might strongly disagree with one of those statements but only somewhat disagree with the other, while in another state, the median respondent might only somewhat disagree with both. The fiftieth percentile would be more prejudiced in the former. Consistent with Becker's theory and thus taste discrimination, they find that the tenth percentile of racial prejudice best predicts the racial wage gap.[2]

[2] They and we ignore the problem that prejudice is measured on an ordinal scale. Their result can also be interpreted as supporting their choice of cardinalization.

In an alternative approach, Glover, Pallais, and Pariente (2017) study a large supermarket chain in France employing significant numbers of North and Sub-Saharan Africans as probationary cashiers. The authors used an implicit attitudes test to measure each manager's bias against North Africans.[3] They find that North Africans were less likely to be offered overtime when assigned to a biased manager. In addition, a given North African worked less rapidly and was absent more frequently when assigned to a biased manager rather than an unbiased manager, providing further evidence of the impacts of manager prejudice on employees.

In the area of criminal justice, a growing literature attempts to identify taste-based discrimination under the assumption that blacks are less prejudiced against blacks than their white counterparts. These studies generally find smaller racial disparities when the decision-maker is black. This has been demonstrated for motor vehicle searches (Anwar and Fang 2006; Antonovics and Knight 2009), automobile crash investigations (West 2018), and jury convictions (Anwar, Bayer, and Hjalmarsson 2012).

Goncalves and Mello (2017) test whether police officers treat white drivers caught speeding more leniently than they do black drivers. Because penalties jump discontinuously at certain thresholds, officers sometimes reduce the penalty by lowering the driver's speed to just under a threshold. The authors show that black drivers were less likely than white drivers to have a reported speed just below the threshold and that this is highly unlikely to reflect differences in true speeding behavior. They also show that fewer than 20 percent of officers account for the racial discrepancy, suggesting that "a few bad apples" drive the racial disparities in police traffic stops.

Statistical Discrimination

The Importance of Information Imperfections in the Theory

Economists have traditionally modeled statistical discrimination as fully rational (Phelps 1972; Arrow 1972a, b); conversely, they have viewed inferences and actions based on false beliefs as a form of prejudice akin to taste discrimination. For example, an employer who inaccurately believes that blacks are less productive than they really are will act much like a Becker-style firm that gets disutility from hiring blacks. We begin this section with an overview of models of statistical discrimination based on differential productivity, self-enforcing disparities, and differential observability. Recently, however, economists have begun to recognize that new information

[3] This particular implicit attitudes test measured the speed with which an individual correctly assigned French or North African names and positive or negative words about worker competence to the right category when competence and French were in the same box (requiring that the same key be typed) and when competence and North African were in the same box. Managers who believe that North Africans are less competent tend to take longer to perform the task in the latter case than in the former.

may correct false beliefs, and so we will then turn to the small literature that models inaccurate statistical discrimination.

Statistical discrimination can arise from true underlying differences between groups in situations where within-group variation is difficult to observe. Suppose that conditional on observable factors, black drivers are more likely to carry contraband. Then, an officer might be much more likely to search the cars of black drivers. Consider an extreme example of a police officer who knows that 5 percent of blacks and 3 percent of whites carry contraband (holding observable variables constant) but cannot distinguish within race who is more likely to transgress. Moreover, say that for this officer (or police department), the threshold for searching is 4 percent: thus, the officer searches all blacks and no whites. Note that differences producing statistical discrimination need not be innate. They may reflect disparities or discrimination elsewhere in the system.

Once disparities have caused statistical discrimination, the outcome can be self-enforcing. In the Coate and Loury (1993) model of self-confirming expectations, also called "rational stereotyping," there can be multiple equilibria, one of which is discriminatory. To gain some intuition, consider a simplified version of their model (from Lang 2007, pp. 277-80). Suppose workers can either invest in themselves (trained) or not (untrained) at some cost. Firms can only observe an imperfect signal of whether the worker is trained that takes on only three values: definitely trained, maybe trained, and definitely not trained. Firms want to assign trained workers to a skilled job and untrained workers to an unskilled job. In a world in which most workers train, a worker with a "maybe" signal probably trained.[4] Firms will assign such workers to the skilled job. In contrast, if few workers trained, someone with a maybe signal probably did not train. Firms will assign such workers to the unskilled job. Depending on parameters, two equilibria can arise with different proportions of workers investing. If whites are in the high-investment equilibrium and blacks in the low, we have a model of discrimination.

In this model of self-confirming expectations, if blacks were convinced that the labor market will reward them if they invest in themselves and employers were convinced that blacks and whites invest in themselves at the same rate, the self-confirming expectations would shift to a new nondiscriminatory equilibrium. In a sense, this conclusion offers some grounds for optimism. Ferguson (1998) argues that schools are often in an equilibrium where teachers have low expectations of their black students, but that it is possible to move to an equilibrium where black students meet the standards of teachers who have been convinced to have higher expectations for them.

In the real world, of course, we generally cannot wave our hands and eliminate discrimination by changing beliefs. Therefore, historical discrepancies due to legal

[4] To keep the presentation simple, we skip the details. In a more detailed description, this statement depends on the probability of trained and untrained workers getting a "maybe" signal. Similarly, later statements in this paragraph depend on the productivity of trained and untrained workers in the two types of jobs.

discrimination are likely to persist. Cavounidis, Lang, and Weinstein (2019) develop a model with two equilibria but in which history matters. In the equilibrium of this model, firms scrutinize their black workers more closely than they do their white workers. Consequently, a larger share of low-performance black workers than of low-performance white workers separates into unemployment. Because productivity is correlated across jobs, the black unemployment pool is more heavily "churned" and therefore weaker than the white unemployment pool. Provided that workers can, to some extent, hide their employment histories, race will serve as an indicator of expected worker productivity. This creates a self-reaffirming dynamic in which it is optimal for firms to scrutinize black workers but not white ones. This model also makes a number of predictions that are consistent with the true state of the world: for example, whites have higher wages on average, but the wage distributions of blacks and whites can overlap; there are shorter unemployment and longer employment durations for whites; and the separation hazard rate into nonemployment will be higher for blacks with low tenure but converge to whites' hazard rate.

Consider another set of assumptions, based on differential observability rather than differential productivity, that could underlie a model of statistical discrimination. Suppose that employers are better at figuring out the productivity of white workers than that of black workers. Then employers will treat black workers more like an average black worker and differentiate more among white workers. In the extreme, employers will pay all blacks the same but pay whites according to their productivity. In this scenario, high-productivity blacks will be disadvantaged relative to whites, but low-productivity blacks will be better off. If blacks and whites are equally productive, on average, they will earn the same wages. From behind a Rawlsian veil of ignorance, a risk-averse person would prefer to be black.

With modifications, this model can produce a black-white wage differential. First, if jobs are differentially sensitive to skill, it will be efficient to place low-productivity workers in jobs that are relatively insensitive to skill and high-productivity workers in jobs where skill is highly valued. With differential observability, white workers earn more because the market does a better job of matching them to jobs. This effect is stronger if skill is multidimensional. If the market cannot tell which blacks should be (say) poets and which should be mathematicians, there will be a larger share of blacks whose skills are mismatched with their jobs, and blacks, on average, will earn less than whites will.

Alternatively, differential observability can affect the incentives of workers to invest in their own human capital. Say that workers can make unobservable investments in themselves (as in Lundberg and Startz 1983). An individual black worker benefits less than a white counterpart does from making unobservable investments, because the black worker is treated more like the average black. Therefore, blacks make fewer unobservable investments. An implication of this model, as Lang and Manove (2011) point out, is that high-ability blacks have a stronger incentive to signal their productivity by making observable investments. This signaling model predicts the surprising fact that blacks get more education than whites with the same test scores in school. However, this model cannot explain the Neal and Johnson

(1996) result, which Lang and Manove confirm, that black men get lower wages when only test scores are used as a control variable. We expect that a model in which blacks have higher observed educational attainment but put in less (unobserved) effort in school might reconcile many of the results in the literature, but this model has not been formalized.

Incorrect Beliefs and Information

If employers believe incorrectly that blacks are less productive than whites, they will behave similarly to employers engaged in taste discrimination. However, models of taste-based discrimination and incorrect statistical discrimination do differ in some implications—like the effect of improving information. This difference has been explored in some experimental studies.

In a public goods experimental game, subjects received a pot of money from which they chose how much to contribute to a public good and how much to retain. The socially efficient outcome in this game requires everyone to contribute everything, but the equilibrium of the static or finitely repeated game is that subjects should hope for others to contribute so that they can act as free riders but not contribute themselves. In the Castillo and Petrie (2010) version of this game, subjects first played with random partners but then learned that they could choose their partners for the remainder of the game. Subjects randomly received one of three treatments: information about the public goods levels in the participants' prior rounds, a photo revealing the race and sex of the other players, and both. In the absence of information, subjects preferred all other race/ethnic groups to blacks even though all groups except whites contributed similarly. However, in the presence of information, there was no impact of race and sex on the ranking of potential partners.

More recently, Bohren, Imas, and Rosenberg (2019) performed an experiment in which subjects made wage offers to potential hires to perform mathematical calculations. In the absence of information on past performance, Indians and males received higher offers than Americans and females, but the male/female pay gap was less than the actual gap in performance, while Americans actually outperformed Indians on the task. When participants learned about the average performance of the different groups and hired additional workers, the offers more closely, but not fully, resembled the actual productivity gap.

There is relatively little nonexperimental research on inaccurate statistical discrimination. Laouenan and Rathelot (2017) show that minorities (African Americans in North America and Arabs, Muslims, and Sub-Saharan Africans in North America and Europe) renting on Airbnb charge substantially less than other Airbnb proprietors do. After controlling for observable characteristics of the rental unit, a small price gap remains. However, minorities benefit more from a measure of the number and quality of reviews. This suggests that the price gap at least partially reflects statistical discrimination. On the other hand, the price gap declines as the number of reviews increases. If renters care only about the expected average review, this is inconsistent with accurate statistical discrimination. The authors conclude that renters engage in inaccurate statistical discrimination. They do not address whether their results can be reconciled with accurate statistical discrimination if renters care about elements of the review distribution other than the mean. This paper demonstrates the difficulty of establishing inaccurate statistical discrimination with observational data. Future research on this topic must make a compelling case that both the pattern of discrimination is inconsistent with taste discrimination and there is no model that rationalizes the observed behavior when information is imperfect but statistically accurate.

Evidence of Statistical Discrimination

There is strong evidence of statistical discrimination in a wide range of settings (for example, on the market for sports cards, see List 2004; on the commercial sex market in Singapore, see Li, Lang, and Leong 2018). This form of discrimination can often be considered acceptable and allows us to make more efficient decisions. For example, many people give up their seat on a bus to someone who appears elderly or pregnant. People presumably reason that, judging statistically based on appearance, these categories of people may benefit more from sitting. In other cases, statistical discrimination may be both undesirable and socially unacceptable. For example, if police stop and search blacks more frequently because they are statistically more likely to be carrying contraband, they will stop many more innocent blacks than innocent whites, and any positive effects from such a policy in reducing crime would need to be balanced against adverse effects both on those stopped and on police/community relations. Statistical discrimination may be socially undesirable even when it is privately beneficial. Each firm may benefit from scrutinizing black workers more carefully, but the effect on total output may be negative, as unlucky black workers spend more time unemployed. And of course, we recognize that statistical arguments may simply obscure prejudice.

The theory of statistical discrimination suggests that providing information about characteristics correlated with race can reduce discrimination. Thus, if blacks are more likely than whites to have been imprisoned for drug offenses, providing information about convictions for past drug offenses may increase employers' willingness to hire black workers. Consistent with this insight, Wozniak (2015) finds that drug testing increased the employment of blacks.

Similarly, firms are less likely to hire workers with known criminal records. Because a higher proportion of blacks have criminal records than whites do, one might expect that preventing employers from inquiring about criminal records, at least at an early stage, would increase black employment. However, if firms cannot ask for information about criminal records, they may rely on correlates of criminal history, including being a young black man. This concern is even greater if employers tend to exaggerate the prevalence of criminal histories among black men, thus leading to inaccurate statistical discrimination. Agan and Starr (2018) investigate "ban the box" legislation in which companies are forbidden from asking job applicants about criminal background. Before such rules took effect, employers interviewed similar proportions of black and white male job applicants without criminal records. Prohibiting firms from requesting this information reduced callbacks of black men relative to otherwise similar whites. Consistent with this, Doleac and Hansen (2016) find that banning the box reduced the employment of low-skill young black men by 3.4 percentage points and low-skill young Hispanic men by 2.3 percentage points. Similarly, occupational licensing increases the share of minority workers in an occupation despite their lower pass rates on such exams (Law and Marks 2009). Prohibiting the use of credit reports in hiring reduced black employment rather than increasing it (Bartik and Nelson 2019).

Taken together, these studies provide strong evidence that statistical discrimination plays an important role in hiring. Additional information, even if it adversely related to being black, reduces reliance on statistical discrimination and can raise black employment. However, this argument does not address how such hiring practices might affect the quality of the pool of workers available for hire. As discussed earlier (in the context of Cavounidis, Lang, and Weinstein 2019), if a set of firms introduces additional information into hiring practices, the quality of the workers they hire increases, but the quality of the pool of workers available to other firms declines. We know virtually nothing about how such policies affect long-run equilibrium.

It has been suggested that algorithms can diminish racial bias in decision-making (Kleinberg et al. 2018). Algorithms use the prior relation between individual characteristics and outcomes to predict outcomes for other individuals. This approach has become particularly popular in criminal justice: courts use algorithms to estimate the risk of future offending, which then informs decisions about bail and sentencing. Of course, algorithms by definition eliminate the risk of human taste-based discrimination. However, if the data used as the basis for the algorithm includes biased outcomes, the algorithm inherits the bias. Thus, if blacks who commit a crime are more likely to be arrested, an algorithm that uses arrest histories to predict recidivism inherits that bias (Mayson 2018). In practice, many predictors in an algorithmic model are correlated with race (zip code, family situation, prior offenses) and together may predict race quite accurately. It should be noted that the direction of this bias can go in either direction: if blacks with a low likelihood of reoffense are more likely to be arrested, a model predicting reoffense could favor blacks (Rambachan and Roth 2019).

In addition, the risk scores generated by algorithms rarely determine the outcome fully. Instead, judges (or other decision-makers) use them to inform their decisions. As judges adjust the recommendation from the risk score, racial discrepancies can increase. For example, imagine a risk score that perfectly estimates a defendant's risk of recidivism. Suppose further that judges, based either on prejudice or the incorrect belief that race has been excluded from the predictors, enforce harsher sentences on black defendants. Then even with the algorithm, the judge's actions will be discriminatory, possibly more than it would be without the algorithm. Overall, recent research on the use of algorithms in practice has found that algorithms do not reduce racial disparities and sometimes increase them: for example, Doleac and Stevenson (2019) look at this issue using sentencing data from Virginia, while Albright (2019) uses data from bail decisions in Kentucky.

Final Thoughts: Discrimination as System

The focus of this essay has been on how economists view discrimination through the prism of the taste-based and statistical discrimination contexts, with an emphasis on the labor market and criminal justice. But discrimination potentially occurs in many domains, including important areas such as housing, education, and medical treatment (for a short summary of the evidence on discrimination in these areas, see Lang and Spitzer forthcoming). Discrimination works as a system, with discrimination in each institution potentially reinforcing disparities and discrimination in other institutions—and with the effects in some cases potentially reaching across generations. Economists, with some exceptions, have tended to ignore or undervalue what sociologists have called the system of discrimination (Reskin 2012, see also the discussion in this issue by Small and Pager) while perhaps doing a better job of recognizing the relation between disparities and discrimination in their models of statistical discrimination.

For example, a key insight of the statistical discrimination literature is that disparities breed discrimination. If blacks are more likely to have been in prison, employers may use race as an indicator of past imprisonment and discriminate against blacks in employment. If discrimination in the justice system makes blacks more likely to have been in prison, discrimination in the justice system causes labor market discrimination. If blacks' weaker labor market performance makes criminal activity more attractive to them, players in the justice system may statistically discriminate against blacks. Looking beyond the domains of the labor market and criminal justice system, discrimination in educational settings can make blacks less prepared to enter the labor force. By creating wage disparity, labor market discrimination may contribute to residential segregation and educational disparities. Earlier discrimination in job markets and housing markets that affects a previous generation of parents creates discrepancies in the quality of public education that children of different races receive, which in turn may create productivity differences and labor market discrepancies across the next generation—even without active discrimination by current employers.

This idea of discrimination as a system is not easy for economists to address. Developing truly general equilibrium models is difficult, especially when the endogenous variables go beyond prices and quantities. Empirical microeconomists, the primary group of economists who study discrimination, have in recent years placed a heavy emphasis on credible identification. While it is possible to imagine studying linkages across the system of discrimination through natural experiments or methods like regression discontinuities and differences-in-differences, it is not trivial to do so. However, the idea of discrimination as a system does suggest some different angles for research and policy analysis.

To the extent that discrimination is a system, efforts to prohibit discrimination in one institution will have only limited effect. Thus, the antidiscrimination policies most likely to be effective will target key leverage points where decreasing discrimination could have strong ripple effects throughout the system (Reskin 2012). In considering policy proposals to address discrimination in the labor market and criminal justice system, we may have to look beyond these two institutions. For example, policies that address discrimination in education can decrease statistical discrimination by decreasing racial disparities among workers entering the labor market.

In addition, policies to increase interracial contact—like limiting residential segregation—may offer a useful point of leverage. Residential and social segregation may lead to prejudice and taste-based discrimination. Pettigrew and Tropp (2006) provide a meta-analysis of 515 studies and conclude that there is strong support for "intergroup contact theory," which proposes that contact tends to reduce prejudice. Some economists have contributed to our understanding of this topic. Carrell, Hoekstra, and West (2015), for example, found that having an additional black member in an Air Force squadron of roughly 35 people increased the probability of having a black roommate as a sophomore (usually not a freshman squadron member) by about one percentage point, or about 18 percent. Similarly, exposure to more black peers with high admissions scores increased the probability that whites reported that they had become more accepting of African Americans. Dahl, Kotsadam, and Rooth (2018) find similar positive effects on male attitudes towards female recruits from having been assigned to a squad with a woman member during boot camp in Norway. In particular, given the importance of networks in job search, social distance can directly increase racial disparities in employment (Loury 2000). These studies, together with the large literature outside economics, suggest a public interest in greater integration and reducing social distance across groups.

References

Abrams, David S., Marianne Bertrand, and Sendhil Mullainathan. 2012. "Do Judges Vary in Their Treatment of Race?" *The Journal of Legal Studies* 41 (2): 347–83.

Agan, Amanda, and Sonja Starr. 2018. "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment." *Quarterly Journal of Economics* 133 (1): 191–235.

Albright, Alex. 2019. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions." Harvard John M. Olin Fellow's Discussion Paper 85.

Alesina, Alberto, and Eliana La Ferrara. 2014. "A Test of Racial Bias in Capital Sentencing." *American Economic Review* 104 (11): 3397–433.

American Civil Liberties Union (ACLU). 2013. *The War on Marijuana in Black and White.* New York: ACLU.

Antonovics, Kate, and Brian G. Knight. 2009. "A New Look at Racial Profiling: Evidence from the Boston Police Department." *Review of Economics and Statistics* 91 (1): 163–77.

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson. 2012. "The Impact of Jury Race in Criminal Trials." *Quarterly Journal of Economics* 127 (2): 1017–55.

Anwar, Shamena, and Hanming Fang. 2006. "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence." *American Economic Review* 96 (1): 127–51.

Anwar, Shamena, and Hanming Fang. 2015 "Testing for Racial Prejudice in the Parole Board Release Process: Theory and Evidence." *Journal of Legal Studies* 44 (1): 1–37.

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics* 133 (4): 1885–932.

Arrow, Kenneth J. 1972a. "Models of Job Discrimination." In *Racial Discrimination in Economic Life*, edited by Anthony H. Pascal, 83–102. Lexington, MA: D.C. Heath.

Arrow, Kenneth J. 1972b. "Some Mathematical Models of Race Discrimination in the Labor Market." In *Racial Discrimination in Economic Life*, edited by Anthony H. Pascal, 187–204. Lexington, MA: D.C. Heath.

Ayres, Ian, Fredrick E. Vars, and Nasser Zakariya. 2005. "To Insure Prejudice: Racial Disparities in Taxicab Tipping." *Yale Law Journal* 114 (7): 1613–674.

Bartik, Alex, and Scott Nelson. 2019. "Deleting a Signal: Evidence from Pre-Employment Credit Checks." MIT Department of Economics Graduate Student Research Paper 16–01; Chicago Booth Research Paper No. 19–23.

Becker, Gary S. (1957) 1971. *The Economics of Discrimination*. 2nd ed. Chicago: Chicago University Press.

Bendick, Marc, Jr. 2007. "Situation Testing for Employment Discrimination in the United States of America." *Horizons Stratégiques* 3 (5): 17–39.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review* 109 (10): 3395–436.

Broyles, Philip, and Bradley Keen. 2010. "Consumer Discrimination in the NBA: An Examination of the Effect of Race on the Value of Basketball Trading Cards." *Social Science Journal* 47 (1): 162–71.

Bryson, Alex, and Arnaud Chevalier. 2015. "Is There a Taste for Racial Discrimination Amongst Employers?" *Labour Economics* 34: 51–63.

Burden, Barry C., Yosikuni Ono, and Masahiro Yamada. 2017. "Reassessing Public Support for a Female President." *Journal of Politics* 79 (3): 1073–8.

Bygren, Magnus. 2010. "The Gender Composition of Workplaces and Men's and Women's Turnover." *European Sociological Review* 26 (2): 193–202.

Carrell, Scott E., Mark Hoekstra, and James E. West. 2015. "The Impact of Intergroup Contact on Racial Attitudes and Revealed Preferences." NBER Working Paper 20940.

Castillo, Marco, and Ragan Petrie. 2010. "Discrimination in the Lab: Does Information Trump Appearance?" *Games and Economic Behavior* 68 (1): 50–9.

Cavounidis, Costas, Kevin Lang, and Russell Weinstein. 2019. "The Boss is Watching: How Monitoring Hurts Blacks." *NBER Working Paper 26319.*

Charles, Kerwin Kofi, and Jonathan Guryan. 2008. "Prejudice and The Economics of Discrimination." *Journal of Political Economy* 116 (5): 773–809.

Coate, Stephen, and Glenn C. Loury. 1993. "Will Affirmative–Action Policies Eliminate Negative Stereotypes?" *American Economic Review* 83 (5): 1220–40.

Combes, Pierre–Philippe, Bruno Decreuse, Morgane Laouénan, and Alain Trannoy. 2016. "Customer Discrimination and Employment Outcomes: Theory and Evidence from the French Labor Market." *Journal of Labor Economics* 34 (1): 107–60.

Dahl, Gordon, Andreas Kotsadam, and Dan–Olof Rooth. 2018. "Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams." NBER Working Paper 24351.

Daly, Mary C., Bart Hobijn, and Joseph H. Pedtke. 2020. "Labor Market Dynamics and Black–White Earnings Gaps." *Economics Letters* 186: Article 108807.

Doleac, Jennifer L., and Benjamin Hansen. 2016. "Does 'Ban the Box' Help or Hurt Low–Skilled Workers? Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden." NBER Working Paper 22469.

Doleac, Jennifer, and Luke C.D. Stein. 2013. "The Visible Hand: Race and Online Market Outcomes." *Economic Journal* 123 (572): F469–F492.

Doleace, Jennifer, and Megan T. Stevenson. 2019. "Algorithmic Assessment in the Hands of Humans." IZA Discussion Paper 12853.

Engel, Robin S., and Rob Tillyer. 2008. "Searching for Equilibrium: The Tenuous Nature of the Outcome Test." *Justice Quarterly* 25 (1): 54–71.

Ferguson, Ronald F. 1998. "Teachers' Perceptions and Expectations and the Black–White Test Score Gap." In *The Black–White Test Score Gap,* edited by Christopher Jencks and Meredith Phillips, 229–72. Washington, DC: Brookings Institution Press.

Fryer, Roland G., Jr. 2019. "An Empirical Analysis of Racial Differences in Police Use of Force." *Journal of Political Economy* 127 (3): 1210–61.

Fryer, Roland G., Jr., and Steven D. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *Quarterly Journal of Economics* 119 (3): 767–805.

Giuliano, Laura, David I. Levine, and Jonathan Leonard. 2009. "Manager Race and the Race of New Hires." *Journal of Labor Economics* 27 (4): 589–631.

Glover, Dylan, Amanda Pallais, William Pariente. 2017. "Discrimination as a Self–Fulfilling Prophecy: Evidence from French Grocery Stores." *Quarterly Journal of Economics* 132 (3): 1219–60.

Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *American Economic Review* 90 (4): 715–41.

Goncalves, Felipe, and Steven Mello. 2017. "A Few Bad Apples? Racial Bias in Policing." Princeton University Industrial Relations Section Working Paper 608.

Grogger, Jeffrey, and Greg Ridgeway. 2006. "Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness." *Journal of the American Statistical Association* 101 (475): 878–87.

Hedegaard, Morten Størling, and Jean–Robert Tyran. 2018. "The Price of Prejudice." *American Economic Journal: Applied Economics* 10 (1): 40–63.

Horrace, William C., and Shawn M. Rohlin. 2016. "How Dark is Dark? Bright Lights, Big City, Racial Profiling." *Review of Economics and Statistics* 98 (2): 226–32.

Jacquemet, Nicolas, and Constantine Yannelis. 2012. "Indiscriminate Discrimination: A Correspondence Test for Ethnic Homophily in the Chicago Labor Market." *Labour Economics* 19 (6): 824–32.

Kahn–Lang, Ariella. 2018. "Missing Black Men? The Impact of Under-Reporting on Estimates of Black Male Labor Market Outcomes." Unpublished. https://scholar.harvard.edu/files/ariellakahn-lang/files/kahn-lang_jmp_20181110.pdf.

Kalinowski, Jesse, Stephen L. Ross, and Matthew B. Ross. 2017. "Endogenous Driving Behavior in Veil of Darkness Tests for Racial Profiling." Human Capital and Economic Opportunity Working Group Working Paper 17.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.

Knowles, John, Nicola Persico, and Petra Todd. 2001. "Racial Bias in Motor Vehicle Searches: Theory and Evidence." *Journal of Political Economy* 109 (1): 203–29.

Knox, Dean, Will Lowe, and Jonathan Mummolo. 2019. "How Administrative Records Mask Racially Biased Policing." Unpublished. https://scholar.princeton.edu/sites/default/files/jmummolo/files/klm_10_2019_w_appendix.pdf.

Lang, Kevin. 2007. *Poverty and Discrimination.* Princeton, NJ: Princeton University Press.

Lang, Kevin, and Jee-Yeon K. Lehmann. 2012. "Racial Discrimination in the Labor Market: Theory and Empirics." *Journal of Economic Literature* 50 (4): 959–1006.

Lang, Kevin, and Michael Manove. 2011. "Education and Labor Market Discrimination." *American Economic Review* 101 (4): 1467–96.

Lang, Kevin, Michael Manove, and William T. Dickens. 2005. "Racial Discrimination in Markets with Announced Wages." *American Economic Review* 95 (4): 1327–40.

Lang, Kevin, and Ariella Kahn–Lang Spitzer. Forthcoming. "How Discrimination and Bias Shape Outcomes." *Future of Children.*

Laouenan, Morgane, and Roland Rathelot. 2017. "Ethnic Discrimination on an Online Marketplace of Vacation Rental." University of Warwick Centre for Competitive Advantage in the Global Economy Working Paper 318.

Law, Marc T., and Mindy S. Marks. 2009. "Effects of Occupational Licensing Laws on Minorities: Evidence from the Progressive Era." *Journal of Law and Economics* 52 (2): 351–66.

Leonard, Jonathan S., David I. Levine, and Laura Giuliano. 2010. "Customer Discrimination." *Review of Economics and Statistics* 92 (3): 670–8.

Li, Huailu, Kevin Lang, and Kaiwen Leong. 2018. "Does Competition Eliminate Discrimination? Evidence from the Commercial Sex Market in Singapore." *Economic Journal* 128 (611): 1570–608.

List, John A. 2004. "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field." *Quarterly Journal of Economics* 119 (1): 49–89.

Loury, Glenn C. 2000. "Who Cares about Racial Inequality." *Journal of Sociology and Social Welfare* 27 (1): 19–52.

Lundberg, Shelly J., and Richard Startz. 1983. "Private Discrimination and Social Intervention in Competitive Labor Markets." *American Economic Review* 73 (3): 340–7.

Mauer, Marc. 2011. "Addressing Racial Disparities in Incarceration." *Prison Journal* 91 (3): 87S–101S.

Mayson, Sandra G. 2018. "Bias in, Bias out." *Yale Law Journal* 128 (8): 2122–473.

McCarthy, Justin. 2019. "Less than Half in the U.S. Would Vote for a Socialist for President." https://news.gallup.com/poll/254120/less-half-vote-socialist-president.aspx.

Mills, Quincy T. 2013. *Cutting along the Color Line: Black Barbers and Barber Shops in America*. Philadelphia, PA: University of Pennsylvania Press.

Mumford, Kevin J. 1997. *Interzones: Black/White Sex Districts in Chicago and New York in the Early Twentieth Century*. New York: Columbia University Press.

Neal, Derek. 2004. "The Measured Black-White Wage Gap among Women is Too Small." *Journal of Political Economy* 112 (S1): S1–S28.

Neal, Derek A., and William R. Johnson. 1996. "The Role of Premarket Factors in Black–White Wage Differences." *Journal of Political Economy* 104 (5): 869–95.

Pettigrew, Thomas F., and Linda R. Tropp. 2006. "A Meta–Analytic Test of Intergroup Contact Theory." *Journal of Personality and Social Psychology* 90 (5): 751–83.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62 (4): 659–61.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett–Davies, Vignesh Ramchandran, Cheryl Phillips, and Sharad Goel. 2017. "A Large–Scale Analysis of Racial Disparities in Police Stops Across the United States." *ArXiv* 1706: Article 05678.

Rambachan, Ashesh, and Jonathan Roth. 2019. "Bias In, Bias Out? Evaluating the Folk Wisdom." *ArXiv* 1909: Article 08518.

Rehavi, M. Marit, and Sonja B. Starr. 2014. "Racial Disparity in Federal Criminal Sentences." *Journal of Political Economy* 122 (6): 1320–54.

Reskin, Barbara. 2012. "The Race Discrimination System." *Annual Review of Sociology* 38: 17–35.

Rosén, Åsa. 1997. "An Equilibrium Search–Matching Model of Discrimination." *European Economic Review* 41 (8): 1589–613.

Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel. 2017. "The Problem of Infra–marginality in Outcome Tests for Discrimination." *Annals of Applied Statistics* 11 (3): 1193–216.

West, Jeremy. 2018. "Racial Bias in Police Investigations." Unpublished. https://people.ucsc.edu/~jwest1/articles/West_RacialBiasPolice.pdf.

Wozniak, Abigail. 2015. "Discrimination and the Effects of Drug Testing on Black Employment." *Review of Economics and Statistics* 97 (3): 548–66.

Zussman, Asaf. 2013. "Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars." *Economic Journal* 123 (572): F433–F468.

Journal of Economic Perspectives—Volume 34, Number 2—Spring 2020—Pages 90–118

Evaluating State and Local Business Incentives

Cailin Slattery and Owen Zidar

Income and opportunity vary substantially across regions. After decades of skepticism (Glaeser and Gottlieb 2008), there is now growing enthusiasm among many policymakers and academics for using place-based policies to address these regional disparities (Austin, Glaeser, and Summers 2018; Bartik 2019b; Kline and Moretti 2014b). Summers (2019), for example, discusses the widely uneven incidence of distress, the inability of natural economic forces and migration to address these disparities, and the political ramifications of growing disaffection of noncosmopolitans. Others emphasize that place-based policies can have unique targeting benefits that transfer resources to distressed regions and can create substantial welfare gains (Gaubert, Kline, and Yagan 2020).

The primary place-based policy in the United States is state and local business tax incentives. Bartik (2019b) estimates that these incentives amount to approximately $46 billion out of $60 billion of local economic development spending and have tripled since the 1990s. Despite the growing enthusiasm for place-based policies in general, many question the effectiveness of these business incentives and whether the mounting costs are justified. Unlike many other federal place-based policies that focus on infrastructure improvements, hiring subsidies, or place-specific tax credits, these business tax incentives are controlled and implemented

■ *Cailin Slattery is Assistant Professor of Economics, Columbia Business School, New York, New York. Owen Zidar is Associate Professor of Economics and Public Affairs, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, New Jersey, and a Faculty Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are cailin.slattery@columbia.edu and ozidar@princeton.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.2.90.

at the state and local level. This autonomy has led some policymakers to propose banning these incentives out of concerns over race-to-the-bottom and beggar-thy-neighbor behavior.

This essay describes and evaluates state and local business tax incentives in the United States. We consider three types of tax incentives—state corporate taxes, state tax credits, and firm-specific incentives—and the trade-offs involved with using them to achieve local and national objectives. Across types of incentives, the key trade-off is between targeting and discretion. Firm-specific incentives can attract marginal firms at lower cost than a corporate tax cut for all firms, but local discretion reduces transparency, and its effectiveness relies on politicians picking winners on the basis of economic rather than political reasons. Across levels of government, local discretion can align with local preferences, technologies, and economic conditions (Oates 1972) but can also result in excessive subsidies, misallocated funds, and negative externalities on other locations (Gordon 1983). Across locations, the key tension is between productive efficiency and equity.

We use new datasets from Slattery (2019) to characterize these incentive policies, to describe the selection process that determines which places and firms give and receive incentives, and then to evaluate the economic consequences. In 2014, states spent between $5 and $216 per capita on incentives for firms in the form of firm-specific subsidies and general tax credits, which mostly target investment, job creation, and research and development. Collectively, these incentives amounted to nearly 40 percent of state corporate tax revenues for the typical state, but in some states, incentive spending exceeded corporate tax revenues. States with higher per capita incentives tend to have higher state corporate tax rates. Recipients of firm-specific incentives are usually large establishments in manufacturing, technology, and high-skilled service industries, and the average discretionary subsidy is $160 million for 1,500 promised jobs. Firms tend to accept subsidy deals from places that are richer, larger, and more urban than the average county, while poor places provide larger incentives and spend more per job.

While we find some evidence of direct employment gains from attracting a firm, we do not find strong evidence that firm-specific tax incentives increase broader economic growth at the state and local level. Although these incentives are often intended to attract and retain high-spillover firms, the evidence on spillovers and productivity effects of incentives appears mixed. As subsidy-giving has become more prevalent, subsidies are no longer as closely tied to firm investment. If subsidy deals do not lead to high spillovers, justifying these incentives requires substantial equity gains, which are also unclear empirically.

The lack of clear spillovers and equity benefits suggests potentially large gains from reforms that direct resources to where efficiency and equity gains are largest. We discuss some of these reforms in the conclusion. Overall, much more work is needed to evaluate the efficiency and distributional consequences of these policies and to design reforms that better address regional disparities and improve the well-being of the unemployed and working class.

Conceptual Framework

Stated Goals of Business Incentive Programs

The stated goal of most state and local business incentives is to stimulate local economic activity, create jobs, and boost wages. For example, the legislation enacting North Carolina's Job Development Investment Grant program states:

> It is the policy of the State of North Carolina to stimulate economic activity and to create new jobs for the citizens of the State by encouraging and promoting the expansion of existing business and industry within the State and by recruiting and attracting new business and industry to the State.

The goals are similar for business incentives that do not explicitly target firm entry and job creation. A report on California's research and development tax credit reads (Hall and Wosinska 1999):

> California is perceived as a high-tax business environment by firms contemplating setting up business or expanding. . . . An R&D-related tax measure targets the particular types of firms that California desires to attract in spite of its relatively high position in the "tax league" tables.

In this section, we provide a framework to consider how state and local governments use different business incentive policies to achieve their objectives of stimulating local economic activity, creating jobs, and boosting wages.

State and Local Objective Function

Consider how business incentives affect local workers, capital owners, and politicians. All else equal, workers benefit from employment opportunities, higher wages, lower local prices, lower taxes, high-quality government services, and other amenities. Capital owners, who include local firm owners and landowners, benefit from higher after-tax-and-incentive profits and rents. Profits depend on intermediate input costs, worker wages, borrowing costs, rental rates of capital, product demand, and productivity, which may be related to government spending on roads and schools. Politicians value improvements in their reelection odds, campaign contributions, pork provision opportunities, and other aspects of their political success. The weights that state and local governments place on the well-being of these different groups may vary across places and over time and may also vary within groups. For example, the weights on the well-being of low- and high-earning residents may differ.

State and local policymakers have several policy instruments—within and outside the tax system—with which to maximize this objective function. We focus on three instruments within the tax system: lowering corporate tax rates, narrowing the corporate tax base, and offering firm-specific tax incentives. Table 1 shows that the average state in 2014 had a corporate tax rate of 6.5 percent, spent $57 per capita on business tax incentives in general, and offered 14 firm-specific tax incentives

Table 1
Business Tax Policy Instruments across States in 2014
(all monetary values in 2017 US dollars)

	Average	*AL*	*CA*	*NV*	*NY*	*PA*	*SC*	*TN*	*WV*
Instrument 1									
Corporate tax rate (%)	6.5	6.5	8.8	0	7.1	10	5	6.5	6.5
Corporate tax revenue per capita	162	90	246	0	264	193	81	193	118
Instrument 2									
Tax credits per capita	19	11	60	0	33	15	32	16	0
Economic development per capita	34	15	2	5	142	25	8	35	177
Instrument 3									
Number of subsidies	14	15	13	4	20	3	16	12	4
Cost per job	45,785	12,466	4,997	42,339	11,712	93,406	6,433	11,805	34,345
Incentives as a percent of corporate tax revenues	38	29	25	NA	66	20	49	26	150

Note: This table shows differences in business tax policy instruments across states in 2014. Corporate income tax revenue is sourced from the US Survey of State and Local Government Finance, via the Tax Policy Center (Urban-Brookings Tax Policy Center 1977–2016). Population data comes from the US Census, and data on corporate tax rates comes from Council of State Governments (1950–2018). The data on tax expenditures, economic development spending, number of subsidies, and cost per job are from Slattery (2019). The number of subsidies and cost per job statistics are available for subsidies that exceed 5 million dollars, as described in the data section. State tax credits per capita are expenditures on tax credits for businesses divided by population, and state economic development per capita is any spending in the state budget for new and expanding businesses, divided by population. The number of subsidies and the average cost per job (Instrument 3) are for the entire sample (2002–2017), not just for 2014. NA is not applicable.

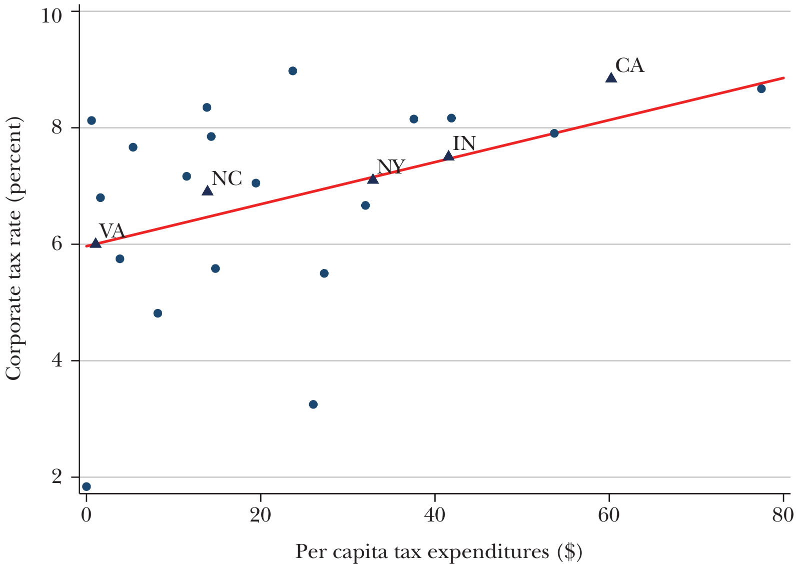
from 2002 to 2017. On average, the general and firm-specific incentives amount to 38 percent of state corporate tax revenue.

There is significant heterogeneity in business taxes and incentives across states. For example, Iowa and Pennsylvania have corporate income tax rates of roughly 10 percent, whereas Nevada and Washington have no corporate tax. State corporate tax apportionment rules and other base provisions vary widely across states as well. Indeed, Suárez Serrato and Zidar (2018) show that state tax base and credit rules explain more of the variation in state tax revenues than state corporate rates do.

Figure 1 shows that states with high corporate tax rates tend to have narrower tax bases (measured as business tax expenditures per capita). States with higher statutory corporate tax rates are able to offer "larger" tax incentives because a tax incentive serves to reduce the effective corporate tax rate. However, a large tax incentive with a high tax rate may still be less attractive than a state with a much lower corporate tax (and no tax incentive), all else equal.

Table 1 highlights the policy differences across eight states. West Virginia has a corporate tax rate of 6.5 percent and corporate tax revenue of $118 per capita, but spends $177 per capita on business tax incentives, all through state

Figure 1
Corporate Tax Rates and Tax Expenditures Are Positively Correlated



Note: This figure summarizes state corporate income tax rates and per capita state business tax expenditures in 2014. The per capita tax expenditure is the total amount of tax expenditures for new and expanding businesses in a state, divided by the state population. The source of the tax expenditure data is the state tax expenditure reports (Slattery 2019). Data on corporate tax rates come from Suárez Serrato and Zidar (2018), and population data come from the US Census. Triangles in the plot are individual data points; circles are binned data. Best fit line estimates are taken from a population-weighted linear regression of corporate tax rates onto per capita tax expenditures.

budget programs. Meanwhile, California has a corporate tax rate of 8.8 percent and double the per capita tax revenue, and offers only $62 per capita in business tax incentives, mainly through tax credits. West Virginia gave only four large firm-specific subsidies between 2002 and 2017, at a cost of $29,000 per job, while California gave 13 subsidies, but at a much lower price of $5,000 per job.[1] What are the costs and benefits of these different approaches, and why might governments adopt different policies?

Local Costs and Benefits of Each Policy Instrument

State and local governments can provide firm-specific incentives in exchange for a firm's commitment to certain levels of investment and employment. Consider

[1] There is also within-state variation over time in how states finance these incentives. For example, in 2013 Michigan decreased per capita corporate tax expenditures from $117 to $13. However, the per capita economic development budget rose by about $90, leaving total per capita corporate incentive spending almost unchanged. The number of subsidies and cost per job are for a sample of large firm-specific deals, which we describe in the section, "Data on State and Local Business Incentives." There are likely more modest subsidy deals with lower cost per job numbers, for which we do not have data because they are not covered by the media.

the welfare of local workers if the government gives a firm-specific tax incentive to a firm that promises 1,000 jobs at $60,000 a year. While some of these jobs go to migrants, local residents who get these jobs will enjoy welfare gains that depend on their prior wages and employment status. The benefits for local workers may extend beyond the directly employed group of 1,000 workers—this new firm may increase local labor demand and therefore local wages (Moretti 2010). However, this entry can also have negative effects. Entry can create congestion and increase land prices. If the tax revenue generated from additional employment and economic activity falls short of paying for the incentive, it will need to be financed through higher taxes or lower public service provision, reducing the welfare of local workers.

The owners of the firm receiving the incentive benefit directly. The owners of other local firms may benefit from productivity and demand spillovers, but might face higher wages, local prices, and taxes. Local landowners will gain to the extent that land prices increase but may be worse off if their property tax bills increase to finance the incentive. Local politicians may benefit from the publicity of a salient plant arrival, increased economic activity, goodwill from recipient firms, and other political considerations.

Instead of lowering the tax bill of a single firm, state governments can lower corporate taxes across the board. Lowering the corporate tax rate mechanically benefits all C corporations in the state by decreasing their tax bill. It also encourages the entry of new firms and the expansion of existing firms (Giroud and Rauh 2019). Like the case of firm-specific tax incentives, this new activity increases the demand for labor and other inputs, which can increase local wages and prices (Suárez Serrato and Zidar 2016). To the extent that lower corporate taxes are financed by reduced public services, a corporate tax cut may have adverse effects on productivity. In Kansas, for example, dramatic business income tax cuts reduced state revenues by $700 million, leading to underfunding of public schools, increases in sales taxes, and decreases in infrastructure spending (Leachman 2017).

For workers, the effects of a corporate tax cut depend on real wages and fiscal conditions. If labor demand increases lead to higher wages and less unemployment, then workers will enjoy some benefit. But if wages and employment don't increase substantially, then workers may be worse off due to higher taxes or lower public service provision. Similarly, the effects on landowners depend on local prices and property taxes.

An intermediate policy intervention often involves narrowing the corporate tax base by lowering the tax bill for a set of firms on the basis of their their activity or industry. For instance, state investment tax credits and accelerated depreciation allowances reduce the tax bill for firms that do a lot of investment (for example, Chirinko and Wilson 2008; Ohrn 2018), and research and development tax credits reduce the tax bill for research-intensive firms (for example, Wilson 2009). This approach has direct effects on recipient firms that are similar to lowering the corporate tax rate, and indirect effects on other firms that resemble those due to firm-specific incentives.[2]

[2] State policymakers also use base provisions to tax different sources of revenue. Through the apportionment system, state policymakers can allocate taxes on the basis of where goods are sold (destination),

Trade-Offs among Policy Instruments

What are the trade-offs among these policy instruments? Consider a typical state that both provides business incentives and taxes corporate income. If providing an additional firm-specific incentive attracts a "high-benefit" firm, the firm's arrival will lead to higher productivity, more prosperity, and higher tax revenue both from other firms and from higher incomes (Bartik 1991; Glaeser 2001; Garcia-Mila and McGuire 2002; Henderson 2003; Greenstone and Moretti 2003; Greenstone, Hornbeck, and Moretti 2010). Offering a firm-specific tax incentive also enables states to contract with firms regarding specific investment and hiring outcomes, which they cannot do with a corporate tax cut.

Firm-specific subsidies can also be used to target mobile firms. If some firms are more responsive to state corporate taxes, then providing tax relief for these firms can allow governments to price discriminate and raise revenues more efficiently (Ramsey 1927).[3] Lastly, firm-specific incentives help attract or retain firms without lowering revenue collected from all firms in the state. A firm that attracts skilled workers, that broadens the industrial mix of an area, or that has hard-to-build relationship-specific capital with local suppliers may be especially valuable (Glaeser, Scheinkman, and Shleifer 1995; Moretti and Wilson 2014; Glaeser et al. 1992; Acemoglu et al. 2016). Base rules share some of these benefits, such as targeting more responsive activity and linking tax benefits to desired outcomes like investment, employment, and innovation (Suárez Serrato and Zidar 2018).

However, implementing firm-specific incentives presents many challenges. The economic rationale for targeting particular firms hinges on the assumption that state and local governments are able to pick winners effectively. Identifying "high-benefit" firms and forecasting a firm's effect on the local economy—including potential agglomeration economies—is a difficult problem for policymakers and academics alike. Moreover, assessing whether the firm would locate elsewhere without the incentive is also hard; firm location decisions are multidimensional and idiosyncratic. There is mixed evidence on what share of firms receiving targeted benefits are inframarginal and thus not influenced by the tax benefit. For example, Bartik (2019a) argues that most deals involve inframarginal firms (Bartik 2018), while using revealed preference approaches, Slattery (2019) finds that deals affect location choice among the largest establishments and Mast (forthcoming) finds limited effects for tax exemptions for mobile firms within New York State. Firm-specific

where goods are produced (source), and where the firm owners live (residence). These options can lead to different effects on local workers, capital owners, and politicians, depending on the structure of production and the responsiveness of multistate firms to tax provisions. In addition, state personal taxes can be an important policy instrument. The rise of pass-through businesses has made state personal income tax rules an increasingly relevant business tax instrument as many private firms (including S corporations and partnerships) face personal income taxation rather than traditional corporate taxation (Smith et al. 2019). Moreover, if workers care about after-tax pay, low personal tax rates may also enable firms to pay workers less, thus attracting firms to low personal tax rate places.

[3] Black and Hoyt (1989) make a related argument about the net fiscal contribution of new firms. If the marginal cost of providing public goods is less than tax revenues generated by the firm, then the government may want to offer the firm subsidies. For a discussion of related considerations regarding international taxation, see Keen (2001) and Dharmapala (2008).

incentives have clear fiscal costs and unclear benefits. Targeting specific firms may not be as effective in practice.

Using firm-specific incentives has additional costs. The lack of transparency in subsidy-giving leaves it more exposed to political capture and noneconomic allocation of funds. An investigation of the Economic Development Agency in New Jersey, for example, found that lobbying and corruption led to inflated subsidy deals (Corasaniti and Haag 2019).

More broadly, providing generous corporate tax incentives requires raising revenue from other taxes. Figure 1 shows high-incentive states have higher tax rates on average. The social cost of higher taxes grows quickly with the size of the tax, and the benefits of incentives diminish with the size of incentives. Thus, providing additional incentives may not always lead to higher welfare in the state, especially at high levels of incentives. In addition, the most distressed places, where the equity gains may be the largest, may not be able to afford to compete for firms with local incentives.

In the following section, to make these considerations more concrete, we examine a tax incentive Volkswagen received in Tennessee. Then we move from anecdotal evidence to more comprehensive data to measure and evaluate how these incentive policies work in practice.

A Firm-Specific Incentive for Volkswagen

In 2008, Volkswagen and the state of Tennessee came to an agreement. Volkswagen (VW) would locate their new assembly plant in Chattanooga, hire 2,000 employees, and spend almost $1 billion. In exchange, VW would receive a discretionary subsidy worth over $500 million.

Many officials championed the deal not only for the 2,000 promised jobs and $1 billion investment but also for the expected indirect job creation and revenue effects. “The Volkswagen investment in this community is going to have a tremendous economic gain for the entire region. I’m confident we’re going to have a very reasonable incentive package when you look at the initial costs of what is being offered compared with a much bigger long-term return,” stated Matt Kisber, the Tennessee commissioner for Economic and Community Development (ECD). He added, “I think Gov. Bredesen and the mayors here are right to treat an assembly plant as worth a large taxpayer investment. There’s nothing quite like the automobile industry to bring in money, raise family incomes and bring in jobs.” Kisber and the ECD projected that in a few years, VW would have an annual payroll of more than $100 million, would help create 14,000 total jobs, and would have a total economic benefit of over $600 million per year (*Chattanooga Times Free Press* 2008).

Business Incentives for Volkswagen

Tennessee used a mix of instruments to attract VW. The $558 million incentive package consisted of state-level tax credits and grants, as well as discretionary tax abatements and in-kind contributions from the state and local government.

Specifically, the deal consisted of property given to VW ($81 million), worker training ($30 million), highway and road construction ($43 million), rail line upgrades ($3.5 million), "enhanced" state job and investment tax credits over 20 years ($200 million), and local property tax abatements over 30 years ($200 million). Moreover, the state promised specialized tax credits for any suppliers that would locate immediately near the VW plant (*Chattanooga Times Free Press* 2008).

Some of these incentives are available to any qualifying firm in the state. For example, a Tennessee company with at least ten new employees and $1 million of investment qualifies for a $4,500 per job credit for one year. These credits become more generous for larger employment and investment levels. On the basis of their projected plans, VW qualified for Tennessee's "enhanced" Jobs Credit, at $5,000 per job over 20 years. Facing a state corporate tax rate of 6.5 percent, VW would be able to use these credits to reduce their tax bill.

The VW subsidy package also included state funds for worker training. More specifically, the state promised at least $12,000 per employee to train each of the 2,000 workers, and to pay for the construction of a technical training center. Some of this money came from Tennessee's "Fast Track" program, which typically provides grants of $4,000 per employee for worker training.[4]

In short, any manufacturing firm entering Tennessee in 2008 would receive incentives from at least three programs: tax relief from the Jobs Credit and Industrial Machinery Credit, as well as grants for job training from the Fast Track program. The size of the firm's investment and the number of jobs at the plant determine the generosity of these incentives.

Of course, the average manufacturing firm is less likely to receive discretionary incentives such as property, infrastructure, and property tax abatements. These discretionary components of a subsidy deal are usually funded through the state budget. Tennessee has an "Economic Development Fund" that can provide additional grant support to companies expanding or locating in the state, but "only used in exceptional cases where the impact of the company on a given community is significant" (Tennesse Department of Economic and Comunity Development 2019).

Lastly, there are capital grants available to "significant projects." Since 2011, this has only included nine companies, including VW, Amazon, GM, and Nissan. According to the Tennessee state budget, in 2008 the Department of Economic and Community Development spent $109 million on business attraction and recruitment. Tennessee's level of incentives is about average across states in 2014, the state spent $16 per capita on tax credits and $35 per capita on economic development programs, while the national average was $19 per capita on tax credits and $34 per capita on economic development (Table 1). However, incentive spending as a

[4] According to the state budget documents (Haslam 2011), Tennessee spent over $53 million on the Fast Track program in 2008 and allocated $71 million for the program in 2009, likely increasing the available funds because of the arrival of VW. Like the tax credits, the size of the Fast Track grant is determined by the company investment, number of new jobs, and wages of new jobs, as well as the types of skills needed and the location of the project. Since 2011, there have been 874 projects in the "Fast Track" program, and firms received about $4,000 per new job.

percentage of corporate tax revenues in Tennessee was only 26 percent, while the national average was 38 percent.

Volkswagen's Location Decision

One reason that Tennessee put together a subsidy package for VW was that the automobile manufacturer was considering many other sites for their new assembly plant. VW's head of overseas manufacturing described the site selection process as follows: "We took 400 Metropolitan Statistical Areas (MSAs) into account. They were narrowed down by different principal criteria to 12 MSAs in seven states. Then the sites that matched our criteria in terms of size and basic infrastructure were researched in detail and analyzed" (Spathelf 2011). VW hired a consulting firm, the Staubach Company, to assist with the site selection process. A team of 25 consultants were employed full time, analyzing the hundreds of potential sites and soliciting proposals from the dozen locations that they short-listed. According to the director of industry-government relations at VW, this location decision was the result of "truly a very close competition," with Chattanooga narrowly beating out a site in Huntsville, Alabama. The Alabama subsidy offer was at least $386 million (Bruns 2008; Bennett 2008).

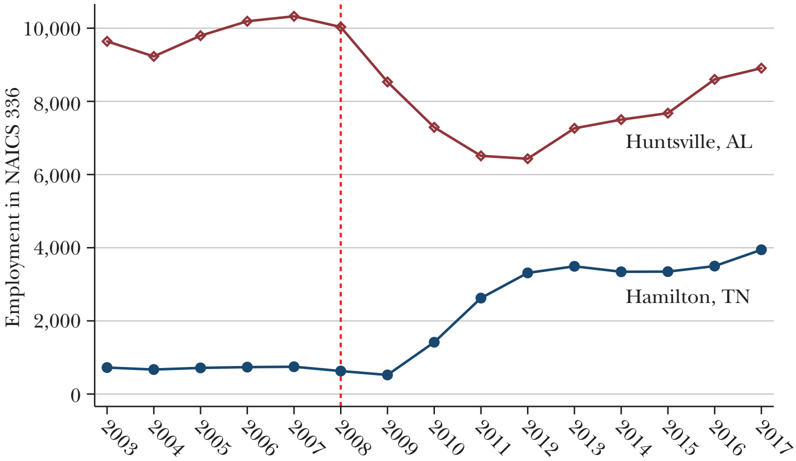
The overall attractiveness of each state depends on more factors than just tax incentives; many nontax considerations matter for VW's location decision. In this case, the two finalists are fairly similar. Both states have right-to-work laws, have state corporate tax rates of 6.5 percent, have similar apportionment weights, and are in the same region of the country—the drive from Chattanooga, Tennessee, to Huntsville, Alabama, takes under two hours. Wages in the sector were higher in Huntsville, and unemployment was slightly lower there as well.[5] Other harder-to-quantify factors, such as the quality of the infrastructure, readiness for a large new assembly plant, and time to build, also influenced VW's decision process.

The Effect of Volkswagen on the Local Economy

We can evaluate this particular subsidy by comparing outcomes in Chattanooga, the winning city in Tennessee, to outcomes in Huntsville, the runner-up in Alabama. Figure 2 plots how employment in Transportation Equipment Manufacturing (NAICS 336, which includes motor vehicle manufacturing as NAICS 3361) evolved from 2000 to 2017 in Hamilton County, Tennessee, and Huntsville, Alabama. The "winning" city, Chattanooga, is in Hamilton County, and the runner-up city, Huntsville, is in Madison County. Huntsville initially had roughly 10,000 employees in transportation equipment manufacturing before the VW deal in 2008, which was substantially more than the roughly 750 employees in Hamilton. After the VW deal in 2008, Hamilton saw a sharp increase in employment to nearly 3,500 employees. The runner-up, Huntsville, experienced a short-term decline of approximately

[5] Transportation manufacturing wages were about $50,000 in Hamilton County and $87,000 in Huntsville. Unemployment was higher in Hamilton at 5.7 percent than Huntsville's 4.3 percent (Bureau of Labor Statistics 1990–2017a). Hamilton County had a population of about 320,000, whereas Huntsville had a population of roughly 400,000. Both locales had per capita income of roughly $40,000. Tennessee had a sales tax rate of 7 percent and no income tax, while Alabama had a sales tax rate of 4 percent and income tax rate of 3.3 percent.

Figure 2
The Employment Effect of Winning Volkswagen



Note: This figure plots total employment in transportation manufacturing (NAICS 336) in Hamilton County, Tennessee, (the county that VW is located in), and Huntsville, Alabama, (the runner-up location in the subsidy competition for VW). The red dashed line indicates the year of the deal, 2008. Data on industry-specific employment come from the Quarterly Census of Employment and Wages (Bureau of Labor Statistics 1990–2017b).

3,500 workers—likely due to the Great Recession—and an eventual recovery back to around 9,000 workers. Thus, the difference in differences amounted to approximately 4,000 workers in transportation manufacturing following the VW deal. If we are concerned that some of this estimate reflects double counting (due to business stealing affecting both the treatment and comparison groups), we can use the raw postdeal versus predeal difference of approximately 2,750 additional jobs in Hamilton.

Recall that the Tennessee commissioner of local economic development projected that VW's promised 2,000-job plant and $1 billion investment would increase local payroll by $100 million, create 14,000 jobs, and have a total economic benefit of $600 million per year. While it is plausible that VW's plant increased auto employment by a few thousand jobs, it is hard to detect effects on total employment.[6] In terms of payroll, the direct estimates were quite reasonable, since 2,000 direct jobs at an average annual salary of $50,000 in 2008 amount to $100 million of payroll. The indirect spillover benefits and "trickle down" of equity impacts are harder to detect.

[6] For example, a simple difference-in-differences specification run on three-digit employment just in these two counties has a treatment effect of 2,679 jobs, but the estimate for total employment is −8,782.

Data on State and Local Business Incentives

Measurement Challenges

The VW deal is a single case study. We need comprehensive data to evaluate these incentives in general. Assembling such data requires addressing some key measurement challenges.

There is substantial complexity and heterogeneity within and across the three broad policies that we consider: reducing the corporate tax rate, narrowing the corporate tax base, and offering firm-specific incentives. Discretionary tax incentives can come with many parts; the subsidy deal for VW was a function of property tax abatements, specialized tax credits, free land, job training, and more. Also, a given firm may qualify for different levels of nondiscretionary tax credits on the basis of its exact industry, investment, and employment, as well as its location within a state and the allocation of activity and sales across states, which affects how multistate firms apportion profits for tax purposes.

An additional measurement challenge is that state and local governments do not usually report the exact amount of tax credits and incentives each establishment in their jurisdiction receives, but instead report tax rates, total tax expenditures, incentive program rules, and incentive program budgets. Because there are many different tax credits and incentive programs, and because the individual tax bill of a firm is unobserved, it is hard to measure and compare incentives across firms and states.

Three Approaches for Measuring Business Incentives

In this subsection we describe three approaches for measuring state and local business incentives: rules-based, expenditure-based, and a narrative-based.

The first approach is "rules-based," which involves collecting data on the rules of each tax, tax credit, and incentive program offered in a locality and predicting the incentives for a firm given its activity and rules. Bartik (2017) applies this approach in the "Panel Database on Incentives and Taxes." This database tracks marginal tax rates and business incentives for 45 industries in 47 cities and 33 states from 1990 to 2015. The focus is incentives for new and expanding businesses. Using the rules of each tax rate, tax credit, and grant in the database, Bartik (2017) uses a simulation model to predict the level of tax incentives a firm would receive in a certain city given the firm's balance sheet.

In terms of state corporate tax rates and tax base provisions, Suárez Serrato and Zidar (2018) also use a rules-based approach. They analyze how rule changes affected state corporate tax revenue over time and across locations, but they are able to analyze only total revenue rather than more granular microdata at the firm or tax-rule level.

The second approach is "expenditure-based," which measures the outlays for each program.[7] Slattery (2019) collects data on state-level business incentives

[7] This approach combines the size of incentives and the level of activity. For example, two states with the same investment tax credit rates could have different investment tax credit expenditures if one state happens to have a lot more (inframarginal) investment.

expenditures from state tax expenditure reports and budget documents by reading each document and identifying tax credits and budget items targeted at new and expanding businesses. The state economic development spending collected from state budgets can include grants, job training, loans, and discretionary subsidies, among other types of incentives. The final product tracks the budget for each program (like job training grants) and expenditure on each tax incentive (for example, investment tax credit) by state and year from 2007 to 2014.

The third approach is "narrative-based," named after the approach for studying national and state tax changes in Romer and Romer (2010) and Giroud and Rauh (2019), respectively. Because systematic reporting on firm-level incentives does not exist, Slattery (2019) uses a variety of sources to assemble a dataset on these firm-level deals. She starts with all subsidy deals worth over $5 million, as reported by the Good Jobs First "Subsidy Tracker" (Good Jobs First 1976–2019).[8] Next, she removes any entry that does not mention expansion, relocation, or a discretionary incentive. Finally, she adds any firm locations reported by *Site Selection* magazine's "Incentives Deal of the Month" columns and annual "Top Deals" reports, arriving at a sample of 543 establishments receiving discretionary subsidies over the period 2002–2017.

Given this list of 543 subsidy deals, Slattery (2019) uses press releases and news articles to fill in details on the number of jobs promised, investment promised, runner-up location, and specific terms of the deal. Each of these subsidy deals combines a range of various incentives, as we saw with VW. The deals often combine state-level tax credits and grants that the state would offer to any company of a certain size, as well as discretionary tax abatements and in-kind contributions from the state and local government. Therefore, in some cases where the news or press release reports only the discretionary component (for example, a discretionary tax credit from the California Competes program), she uses state-level spending data to determine what nondiscretionary incentives are available to the firm (like California's research and development tax credit) and thus to estimate the tax credit for the subsidy deal.[9]

[8] The mission statement of Good Jobs First is "Tracking Subsidies, Promoting Accountability in Economic Development." To this end, they publish a "Subsidy Tracker" that lists all the firm-level incentives they come across. The sources of these incentives include state and local government reports, newspaper articles, press releases, and the results of Freedom of Information Act requests to state governments. There is a lot of selection in the set of subsidies observed in the data, due to differences in reporting across states or even within states but across programs. For the largest subsidy deals, where there is press available, Good Jobs First has much better coverage. Another note is that there are often duplicates in the data, due to subsidies being reported in press releases and tax expenditures being reported annually for states that report tax expenditures at the firm level. The total value of tax incentives may also be underreported for states like California, where the "Subsidy Tracker" has entries for the "California Competes" discretionary tax credit program but does not report how much firms receive from California's generous research and development tax credit.

[9] Over 30 percent of the subsidy deals in our sample mention contributions to the subsidy package from local governments, like county and city governments. This estimate is likely a lower bound on deals that involve local government spending. We do not have a comprehensive dataset on spending at the more local level. We suspect that the local contribution is reported in news articles and press releases when it is a significant portion of the total deal. Like in the VW case, local governments usually add to the subsidy deal by offering property tax abatements, which can be very large in localities with high property taxes. Larger cities may have economic development offices and economic development teams of their own, which will work with the state to develop a subsidy offer for a given firm.

The Magnitude of State and Local Business Tax Incentives

Using a rules-based approach, Bartik (2017) estimates that state and local governments spent $45 billion on incentives in 2015—of which $13.5 billion is attributed to local property tax abatements. After accounting for all local incentives, estimated total state incentive expenditure less than $30 billion. Using an expenditure-based approach, Slattery (2019) estimates that state governments spent about $20 billion on incentives in 2014. Assuming the same relative contribution from local governments as the Bartik (2017) estimates would imply that total state plus local incentives amount to $30 billion in 2014.

Differing approaches and data coverage explain some of the gap between these estimates. Bartik (2017) assumes each firm receives all of the incentives it is qualified for and calculates the level using a simulation model, whereas Slattery (2019) records the state-reported budget for each incentive program and expenditure for each tax credit from the annual budget documents and tax expenditure reports. The state-level dataset of Slattery (2019) does not include local incentives, which are included in Bartik (2017), and is based on reported spending instead of simulated spending given rules and activity. Therefore, the state-level data from Slattery (2019) that we present is a lower bound for total state and local business incentives.

Previous efforts to collect data on firm-specific incentives using a narrative-based approach have found estimates of business incentive spending that are almost twice as high. Main sources for these estimates are the nonprofit Good Jobs First and journalists at the *New York Times* (Storey, Fehr, and Watkins 2012). The *New York Times* sourced data from Good Jobs First's "Subsidy Tracker," consultants, state agencies, and government reporting to create their "United States of Subsidies" interactive database on business incentives. The resulting database reported total business incentive spending of $80 billion in 2012. One reason for this difference from Bartik (2017) and Slattery (2019) is double counting. Firm-specific subsidies can be recorded twice, both as the individual subsidy deal and as part of the total state spending on an incentive program. Also, Bartik (2017) and Slattery (2019) explicitly focus on incentives for new and expanding business, whereas the *New York Times* data include other components such as sales tax exemptions that could apply to individuals and existing companies.

In the Slattery (2019) firm-level incentive data that we use, we take the size of the subsidy given, even though there may be differences in reporting across localities and even industries in how much a given package of incentives is "worth" to a firm. Furthermore, the subsidy size is normalized to 2017 dollars and a ten-year contract. The modal subsidy deal is paid out over ten years, but some have longer horizons, such as the 20 years of state tax credits for VW.

With 543 firm-specific subsidies over the period of 2002–2017, the total amount of discretionary incentives promised was $82 billion, or about $5 billion a year. Subsidy-giving fluctuates over time but generally increased from 2002 to 2014, with a minimum of 14 discretionary incentives offered in 2003 and a maximum of 53 discretionary incentives offered in 2012.

We use per capita incentive spending—the sum of economic development and tax expenditures—as a measure of the generosity of a state's incentives. Top per capita spenders include Michigan, West Virginia, New York, Vermont, and New Hampshire. Their per capita incentive spending in 2014 amounts to 56 percent of public safety expenditures, 40 percent of spending on health and hospitals, 30 percent of transportation, and 12 percent of education. For the full sample of states, it is 23 percent of public safety, 13 percent of health and hospitals, 11 percent of transportation, and less than 5 percent of education. Compared with state corporate tax revenues, incentive spending is about 40 percent of corporate tax revenues on average across states in 2014. There are five states—Nevada, South Dakota, Texas, Washington, and Wyoming—that have zero corporate income tax revenue but spend about $44 per capita on incentives for firms. One measure of generosity of these incentives is the ratio of subsidy to promised jobs, or "cost per job." Interpreting this measure requires care because the subsidy is a flow over a period of ten years, so cost per job *per year* is lower than these estimates. Discount rates and job churn complicate estimates of cost per effective annual full-time employees, but dividing by ten provides a crude, optimistic estimate of cost per job per year. We find that average cost per job has increased over the same period. This finding is consistent with Bartik (2017), who shows that incentives have increased over a longer time period—incentive spending as a percentage of gross taxes increased from 10 percent in 1990 to 30 percent in 2015.

At the time, the VW location deal was the largest subsidy offer made by Tennessee—$558 million for a 2,000-job automobile plant, with a cost of about $279,000 per job promised. However, in terms of discretionary subsidies offered to large firms, it is not an extreme outlier. Over the entire sample of discretionary incentives, firms receive $160 million on average and promise about 1,500 jobs at the establishment. The effective cost per job is $108,000 at the mean but varies a lot over deals—from $11,500 at the tenth percentile to $858,000 at the ninetieth percentile. In the next section, we explore how differences across industries and locations help explain variation in observed subsidy size.

The Allocation of State and Local Business Tax Incentives

This section uses the Slattery (2019) data to assess how well the allocation of incentives aligns with the efficiency, equity, and political goals of policymakers. We first consider how well the allocation of incentives aligns with employment, growth, and spillover goals and then consider how the allocation of incentives aligns with equity and political goals.

Large, profitable firms are more likely to receive firm-specific subsidies. Comparing subsidy receipt by establishment size to the universe of establishment entry in the Census Business Dynamics Statistics reveals that more than 30 percent of all establishments with over 1,000 employees receive discretionary subsidies, while the percentage is less than 0.2 percent for establishments with under 250 employees. The firms that receive discretionary subsidies not only have larger establishments, but they are also larger than the average public company in terms of employment, profits, revenue, and capital stock.[10] The differences are striking. Firms that receive discretionary subsidies from states have eight times as many employees as the average firm in Compustat (60 times more at the median). The gross profit of the average firm in Compustat from 2001 to 2014 is just over $1 billion. The average gross profit for the subsample of firms that ever received a discretionary subsidy in that period is $14 billion, and it is even higher in the year of the subsidy deal ($21 billion).

The fact that larger establishments, which are part of large and profitable firms, are most likely to receive discretionary subsidies is consistent with the following hypotheses on subsidy-giving: state and local jurisdictions try to attract large, productive establishments that will generate surplus and spillovers, affect the location decisions of other establishments, and increase demand for both labor and services. However, the fact that only the largest firms receive large discretionary subsidies may facilitate increasing industry concentration—a topic of much interest and concern (see, for example, DeLoecker, Eeckhout, and Unger 2019).

The prevalence of firm-specific subsidies in manufacturing, technology, and high-skilled services also appears consistent with stated objectives. These industries represent nearly half (47 percent) of our sample of deals. Table 2 shows mean deal characteristics for five industries. Automobile manufacturing firms are the most "popular" industry, with 56 subsidies, or 10 percent of the total sample. The average automobile manufacturer promises to create almost 3,000 jobs and receives $260 million at over $90,000 per job. The Economic Policy Institute estimates that the auto industry has the largest jobs multiplier, with 14 jobs created in the local economy for every one job created at an automobile manufacturing plant (Bivens 2019). Therefore, the prevalence of auto subsidies is consistent with the hypothesis that policymakers target firms with large agglomeration effects.

Other industries get large subsidies, but do not promise as many jobs. For example, basic chemical manufacturing subsidies amount to over $900,000 per job. Establishments in this industry, however, often make large capital investments, have high fixed costs to entry, reduce local energy costs for other firms, and become settled future taxpayers. They may also lobby aggressively. In Louisiana, for example, where many chemical manufacturers are located, a proposal to restrict the discretionary tax exemption program in the state was met with a rush of activity from the chemical industry, and the publication of industry-backed studies on the positive impact that they have in the state (Karlin 2018; Crisp 2016). An open question is how subsidy receipt depends on the product and labor market characteristics of firms as well as the number and type of workers it attracts.

[10] We compare the size of establishments in the subsidy data with the size distribution of establishments entering the United States, from the Census Business Dynamics Statistics (online Appendix Table A.1 available with this article at the *Journal of Economic Perspectives* website). We match the firms in our sample of subsidized firms to firms in the Compustat database (online Appendix Table A.2). Our subsidy dataset is selected on subsidies that are at least $5 million, so smaller firms are likely also receiving smaller discretionary subsidies. However, when we use *all* establishment-level tax credit and grant receipts from the state of Indiana—a state that provides more comprehensive data—we find a very similar pattern on establishment size.

Table 2
Select Industries Receiving Firm-Specific Subsidies

	Number of deals	*Subsidy* (2017 M USD)	*Jobs promised*	*Cost per job (2017 USD)*	*Investment promise* (2017 M USD)
Full sample (4-digit industry codes)	543	178.4	1,487	119,972	757.5
Automobile manufacturers (3361)	56	293.6	2,768	106,057	854.8
Aerospace manufacturers (3364)	31	585.8	2,734	214,237	534.5
Financial activities (5239)	25	92.3	2,652	34,809	286.8
Scientific R&D services (5417)	22	113.7	518	219,259	185.0
Basic chemical manufacturers (3251)	18	187.4	196	956,701	779.0

Note: This table shows descriptive statistics for firm-specific incentives at the industry level, for select industries mentioned in the text. It is tabulated using firm-level subsidy data from Slattery (2019); the sample is firm-specific incentive deals from 2002 to 2017. This sample focuses on large subsidy deals that exceed $5 million, as described in the section, "Data on State and Local Business Incentives." The table reports the number of deals for each industry, mean subsidy size, jobs promised, cost per job, and investment promised. See online Appendix Table A.3 (available with this article at the *Journal of Economic Perspectives* website) for the mean and median characteristics for the top ten industries, by subsidy receipt.

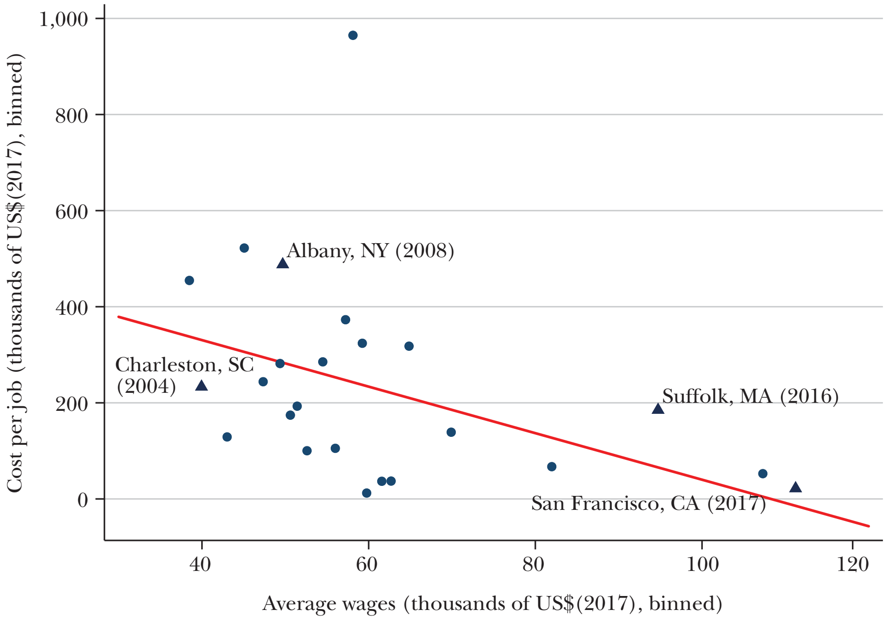
At the state level, we find similar patterns for nondiscretionary incentives. The most popular tax credits (in terms of both the number of credits available and total spending) target job creation, investment, and research activity. Those three types of credits make up 75 percent of total per capita tax expenditures.

Distressed places may enjoy larger welfare gains from increasing local economic activity. Consistent with this observation, we find that within the sample of winning counties, poorer counties are more likely to give larger subsidies. Figure 3 illustrates this phenomenon with a binned scatterplot, with subsidy per job plotted against average wages in the county. It shows that counties with an average wage of less than $40,000 pay over $350,000 per job in the mean subsidy deal. Meanwhile, counties with average wages over $100,000 pay less than $100,000 per job in a given subsidy. This pattern may also be driven by differences in profitability—distressed places may need to provide larger incentives to attract firms. There is little evidence on who benefits from these policies across the income distribution and whether those experiencing wage gains were mostly prior residents, unemployed, or working-class individuals. These are open and important questions.

However, the most distressed locations are rarely able to attract firms with subsidies. Winning and runner-up counties have higher per capita income and higher average wages than the average county.[11] The personal income per capita

[11] In 2000, the winning counties had a mean population of 407,000 and an average wage of $45,500, while the runner-up counties had a mean population of 610,000 and an average wage of $48,600. Counties with at least 100,000 residents have a similar profile to the winners: a mean population of 400,000 and an average wage bill of $44,400. Meanwhile, the average county had a population of 91,000 and an average wage of $34,800. See online Appendix Table A.4 for more detail.

Figure 3
Low-Wage Locations Provide More Generous Subsidies



Note: This figure shows the relationship between local average earnings and the generosity of firm-specific incentives subsidies. The sample is firm-specific incentive deal winners 2002–2017. We calculate the cost per job for each subsidy in the firm-specific deal data (Slattery 2019). Cost per job is the total subsidy size divided by the number of jobs promised in the subsidy deal. Average wages are sourced from the Quarterly Census of Employment and Wages (Bureau of Labor Statistics 1990–2017b). The figure plots total subsidies for each deal relative to average wages in the winning county. Triangles in the plot are individual data points; circles are binned data. Best-fit line estimates is taken from a population-weighted linear regression of subsidy cost per job on average county wages.

of the average county is $34,000, which is $7,000 lower than winning counties and more than $10,000 lower than the runners-up. Facing tight budgets, the most distressed places where equity gains may be largest are not well served by locally financed incentive policies.

We compare the relative importance of determinants of incentive provision at the state level and find a strong role for political factors. In particular, we follow Slattery (2019) by estimating a linear probability model in which the dependent variable equals one if the state increased per capita incentive spending by at least 20 percent. The explanatory variables of interest include whether it was an election year, state GDP per capita in the previous year, state employment rate in the previous year, and whether the governor of the state could run as an incumbent. This specification includes state and year fixed effects. Table 3 presents the results. We find that a 20 percent increase in state per capita incentive spending is less likely when previous-year employment is higher. When a state loses jobs, the fiscal externality of creating a new job is higher. While these economic determinants are important, political factors are especially important. The interaction between

Table 3
When Do States Increase Incentive Spending?

	Per capita incentives increase by 20 percent					
Governor can run as incumbent	0.05				−0.02	−0.02
	(0.06)				(0.06)	(0.06)
Election year		0.11*			−0.08	−0.07
		(0.06)			(0.10)	(0.10)
GDP per capita ($1,000) in $t-1$			0.00			0.02*
			(0.01)			(0.01)
Percent of population employed in $t-1$				−0.05		−0.09**
				(0.03)		(0.04)
Governor can run as incumbent × election year					0.27**	0.25**
					(0.11)	(0.11)
Observations	336	336	336	336	336	336
R^2	0.17	0.18	0.17	0.18	0.20	0.21

Note: This table shows the relationship between state characteristics and increases in state per capita incentive spending. Sample is states between years 2007 and 2014. State per capita incentive spending includes both state tax expenditures on tax credits for businesses and state economic development programs for businesses. The dependent variable is an indicator for whether per capita incentive spending increased by more than 20 percent. States increased per capita spending by over 20 percent 63 times, or 19 percent of the sample of state-years. GDP is sourced from the US Bureau of Economic Analysis (1967–2017). Population is sourced from the 2000 US Census, while employment is sourced from the Census County Business Patterns (US Census Bureau 1997–2017). Data on whether the governor can run as an incumbent or whether the state is in an election year are sourced from Follow the Money (2000–2016). State and year fixed effects are included in each specification. Standard errors are reported in parentheses.
Note: *, **, and *** indicate significance levels of $p < 0.10$, $p < 0.05$, and $p < 0.01$, respectively.

an incumbent governor and an election year is highly correlated with increases in incentive spending, suggesting a strong role for political determinants of incentive provision. In the raw data, per capita incentive spending increases by more than 20 percent in half of the cases in which it is an election year and the governor is up for reelection versus one-fifth of the cases otherwise. This result is consistent with Jensen and Malesky (2018), who find that offering incentives increases the reelection odds of governors.

Effects of State and Local Business Incentives

What is the effect of these business tax policies on firm location, economic activity, and fiscal outcomes? Answering this question is difficult. One approach is to measure how outcomes change after a business tax policy change, but this approach is problematic because state and local governments may be more likely to enact policies or give more generous incentives when economic conditions are deteriorating or are expected to deteriorate. Indeed, in terms of generosity, Figure 3 shows that poorer places give larger subsidies per worker. Alternatively, places may be more likely to enact policies when economic conditions give them more slack in the budget. Thus, comparing outcomes before and after may reflect how the place was trending, rather than the effect of the policy itself. An alternative comparison group, such as runner-up locations or states that will adopt similar policies a few years later, can help difference out common trends and identify the effects of policy changes.

Evidence on Discretionary Subsidies

This subsection investigates the effects of firm-specific tax incentives by comparing outcomes in "winning" locations to "runner-up" locations as in Greenstone and Moretti (2003) and Greenstone, Hornbeck, and Moretti (2010). We examine effects on employment within the targeted industry (at the NAICS three-digit level), spillovers to other labor market outcomes, and overall effects on house prices. We use county-level data to look at these outcomes and state-level data to look at effects on government expenditures and revenues. We then compare these results with prior results in the literature.

Figure 4, panel A, shows how employment within the three-digit industry of the targeted firm differs between winning and runner-up counties in the years before and after the incentive. It shows that differences in the years before the policy are fairly stable but then increase after the discretionary incentive to be approximately 1,500 jobs higher in the winner versus the runner-up location. This result suggests that we can detect the direct effects of these policies on local employment within the sector of the deal.

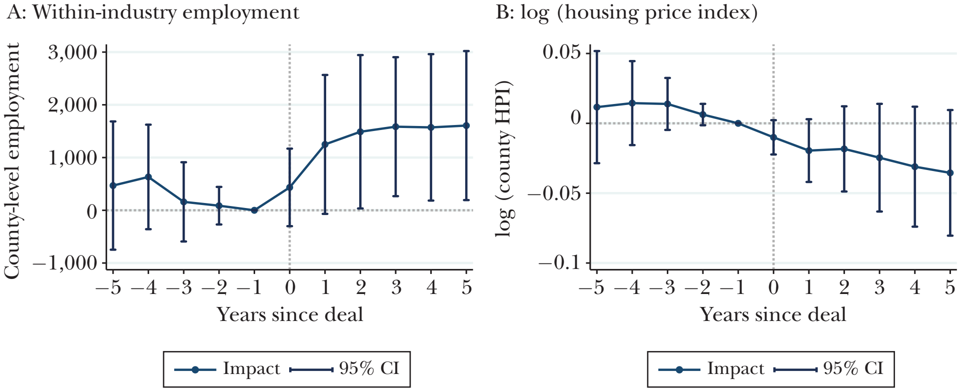
Table 4 shows difference-in-differences estimates for a broader set of local outcomes. The results are more mixed and weaker for effects on employment outside the three-digit industry as well as on countywide outcomes. For example, we don't see strong evidence of effects on other two-digit, one-digit, or countywide employment outcomes outside the directly affected three-digit sector. Patterns are similar at a broader geographic (Consistent Public Use Microdata Areas) level. While it is hard to measure distributional outcomes using publicly available data, column 7 also finds little impact on the employment-to-population ratio, which is one of the ways to measure equity impacts.

Figure 4, panel B, shows the effects on county-level house prices. In the years prior to the discretionary subsidy, winning counties had slightly higher house prices, but the differences are minor. However, after the subsidy, house prices seem to decline to about 4 percent lower relative to the runner-up locations, but the effect is marginally significant statistically. This apparent decline in house prices, which may reflect a negative pretrend rather than real effects, provides some weakly suggestive evidence that the welfare effects of these deals might be negative on average.

Prior studies have examined the effects of firm-specific incentives using the same approach and found somewhat different results. Greenstone and Moretti (2003) found employment effects at the one-digit level and small increases of 1.1–1.7 percent in property values in a sample of 82 primarily manufacturing deals in the 1980s and early 1990s.[12] Greenstone, Hornbeck, and Moretti (2010) use data on productivity

[12] Note, however, that their property value measure differs from the house price index and in their data appendix, they note a 0.54 correlation between their measure and a repeat sale house price index from Office of Federal Housing Enterprise Oversight.

Figure 4
The Effect of Winning a Firm on County-Level Outcomes


A: Within-industry employment

B: log (housing price index)

Note: This figure shows the event study estimates of the effect of winning a firm-specific deal on county-level employment within the NAICS 3-digit industry of deal (Panel A) and on county-level house prices (Panel B). The sample is a balanced panel of firm-specific deals taking place between 2002 and 2012 taken from Slattery (2019). The sample is restricted to deals for which we observe employment, population, and average wages ten years before the year of deal and five years after. Housing price index (HPI) data is taken from the Federal Housing Finance Agency (2014), and employment data is taken from the Quarterly Census of Employment and Wages (Bureau of Labor Statistics 1990–2017b).

Table 4
The Effect of Winning a Firm-Specific Deal on County-Level Outcomes

	3D industry employment (1)	*Residual: 2D industry employment (2)*	*Residual: 1D industry employment (3)*	*Residual: Countywide employment (4)*	*Personal income (5)*	*log HPI (6)*	*Emp/pop (7)*
A. Levels estimates							
Winner × post	1,108.287**	780.238	53.154	−1,920.430	−1,090.989	N/A	−0.001
	(539.686)	(1,096.283)	(1,928.740)	(5,301.175)	(716.305)	N/A	(0.002)
Mean of outcome	9,326.605	15,763.784	49,393.076	2.80e+05	49,826.006	N/A	0.470
B. logs estimates							
Winner × post	0.149**	0.026	0.030	0.003	−0.005	−0.040*	−0.002
	(0.068)	(0.027)	(0.019)	(0.013)	(0.012)	(0.021)	(0.004)
Mean of outcome	7.965	9.037	9.922	12.006	16.667	4.858	−0.759

Note: This table shows difference-in-differences estimates of the effects of winning a firm-specific deal on a variety of county-level outcomes. Employment data are from the Quarterly Census of Employment and Workers (Bureau of Labor Statistics 1990–2017b), HPI data are from the Federal Housing Finance Agency (2014), and employment-to-population figures are computed using Bureau of Labor Statistics data for the numerator and Bureau of Economic Analysis data for the denominator. Standard errors are reported in parentheses and are clustered at the state level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

of other manufacturing plants and find quite substantial total factor productivity spillovers in winning areas. Patrick (2016) finds that the specific estimates in Greenstone, Hornbeck, and Moretti (2010) are somewhat sensitive to research design and specification choices such as the inclusion of trends and the selection of runners-up, and that attracting a new, large plant leads to modest increases in economic activity. Bloom et al. (2019) expand the million-dollar plant dataset and find spillovers using establishment-level data for their sample in the 2000s.

The difference in results across these studies is likely due to differences in the question being asked and how the analysis samples are constructed. Asking what happens in places that had large investments is different than asking what happens in places that gave large subsidies. Large investments may lead to spillovers, whereas large subsidies may not.[13] The Greenstone and Moretti (2003), Greenstone, Hornbeck, and Moretti (2010), and Bloom et al. (2019) sample is selected on the size of firm investment, while our data is selected on the size of the subsidy. As subsidy-giving has become more prevalent in the 2000s, it is not always for large investment projects.[14]

Overall, our assessment of the current evidence is that the spillover effects of subsidies are mixed. While the million-dollar plant papers find evidence of spillovers, large firms that often get large subsides may not generate large spillovers on average in our data and in some other recent work. Criscuolo et al. (2019), for example, examine a related policy in the United Kingdom, where firms apply for a discretionary grant. They find employment and wage effects, but don't see any effects on total factor productivity. Their effects are concentrated for the smallest firms, as larger firms are able to receive incentives without changing their behavior.

Evidence on Corporate Tax Rate and Base Changes

There is also a large literature on the effect of state corporate tax policies on firm location, economic activity, and fiscal outcomes. A number of recent papers show that US state corporate taxes affect firm location and foreign direct investment.[15] In terms of distributional effect, most of the gains from state corporate tax cuts go to firm owners and landowners (Suárez Serrato and Zidar 2016); while there are also some wage gains for workers on average, it is not yet well established

[13] In a recent working paper using Census microdata, Patrick and Partridge (2019) find evidence of small spillovers on incumbent plans in the million-dollar plant sample, and smaller to no spillovers in a larger sample of subsidized plants. The larger sample of subsidized firms is the million-dollar plant sample plus the *Site Selection* sample without runners-up plus the Good Jobs First sample. More generally, one potential issue with the runner-up design is that some firms may use implausible runner-up locations to game the bidding process, in which case effects may be overstated.

[14] To confirm that our specification choices were not causing the differences, we estimated them using the set of plants from Greenstone and Moretti (2003); Greenstone, Hornbeck, and Moretti (2010), and Bloom et al. (2019) and confirmed that our approach delivers results that are consistent with this earlier work in that we see spillovers beyond the three-digit industry at the county level. We report these results in the online Appendix.

[15] Readers interested in this literature might begin with papers from the last few years like Ljungqvist and Smolyansky (2014); Suárez Serrato and Zidar (2016); Fuest, Peichl, and Siegloch (2018); Curtis and Decker (2018); Fajgelbaum et al. (2018); Giroud and Rauh (2019); and Mast (forthcoming)—and then work their way back into the earlier literature cited in those papers. For foreign direct investment and state taxes, Hines (1996) is a central reference.

which workers benefit and how much these benefits "trickle down" to low-wage and unemployed residents.

Some of the more important corporate tax base provisions are those related to innovation and investment. Wilson (2009), Moretti and Wilson (2014, 2017), and Akcigit et al. (2018) study the effects of state tax policies related to research and development on innovation. Ohrn (2018) shows the effects of accelerated depreciation policies on investment at the state level, and Chirinko and Wilson (2008) study state investment tax credits. Garrett, Ohrn, and Suárez Serrato (forthcoming) show that these investment effects also affect state employment. A full assessment of this literature is beyond the scope of this paper, but we mention this research to highlight some of the current evidence on numerous levers policymakers use to attract firms and increase economic activity in their jurisdictions.

National Welfare and Additional Considerations

In practice, firm-specific incentives represent the large majority of local and economic development spending in the United States (Bartik 2019b). From a national perspective, we need to consider not only the effects on the agents in the location providing incentives, but also on the agents in all other locations. Much of the literature on tax competition (for a survey, see Agrawal, Hoyt, and Wilson 2019) highlights the possibility of a race to the bottom and over-subsidization of firms. A prisoner's dilemma perspective in which every location acting in its own self-interest, leading to a suboptimal equilibrium for all, helps reveal why some policymakers have called for subsidy bans in the United States (Markell 2017) and has led to bans on such subsidies within the European Union. In this section, we discuss the conditions under which allowing state and local business incentives may or may not be in the national interest.

As a baseline, consider the simplest frictionless benchmark with no externalities and the optimal level of government service provision. In this case, business incentives likely reduce aggregate welfare. Moving a firm to a location to which it would not go in the absence of incentives misallocates resources (Gaubert 2018; Fajgelbaum et al. 2019) and has fiscal costs that either raise taxes, which increase deadweight loss, or lower government spending below optimal levels.

Considering fiscal externalities and productivity spillovers can change this assessment. If firms do not internalize the externalities they provide, allowing local areas to align tax incentives and social benefits may increase allocative efficiency (Glaeser 2001; Ossa 2015). For example, moving a technology firm from San Francisco, California, to Columbus, Ohio, may lead to more service employment, wage growth, and local fiscal benefits in Ohio. In terms of externalities across locations, reallocating activity may affect fiscal conditions in other locations (Gordon 1983) and can also increase overall economic activity in some circumstances. For example, overall US innovative activity may be higher if the technology firm remains in San Francisco (Moretti 2019; Glaeser and Hausman 2019; Sollaci 2019) due to gains from concentrating scientists in knowledge hubs. It is not clear, however, whether concentrating firms and workers would increase aggregate activity in most industries or in other settings.[16]

Equity considerations are the basis for some of the most compelling arguments for place-based incentives. Income and opportunity vary substantially across regions, and place-based policies can provide unique targeting benefits for addressing these disparities. Gaubert, Kline, and Yagan (2020) characterize these conditions, which relate to sorting, productivity differences across locations, worker mobility, and other features affecting the equity and efficiency of the optimal income tax system.

Moreover, regional disparities may reflect labor market frictions or other distortions from the tax system (Albouy 2009; Fajgelbaum et al. 2019), transfers (Baicker, Clemens, and Singhal 2012), and other state and local policies and regulations (Hsieh and Moretti 2019). For example, search frictions can lead to large gains from incentivizing firms to move to high unemployment areas (Bilal 2019). The underlying idea is that the shadow value of a job or resources in general is likely highly unequal across regions. In principle, the theory of second-best suggests that incentive policies may be able to improve welfare because we are not starting from an undistorted frictionless benchmark. But it is also not clear how effective these policies are in increasing income and opportunity for unemployed and low-wage workers who play a key role in these equity arguments.

The equity gains from local business tax incentives may be limited if the most distressed places, which may benefit most from attracting a firm, lack sufficient revenue to offer incentives to attract that firm. Other prominent place-based poli cies—like the Tennessee Valley Authority discussed by Kline and Moretti (2014a) and Empowerment Zones discussed by Busso, Gregory, and Kline (2013a)—avoid these limitations by using federal funding.

Conclusion

State and local governments are devoting substantial resources toward attracting firms and corporate capital. This article has discussed three different incentive policies state and local governments currently use, the trade-offs between these policies, and the evidence we have on the effects on local economic activity.

We provide descriptive evidence that industries with larger multiplier effects are more likely to receive subsidies and receive more subsidy dollars per job. We also find that poorer places spend more per job. In terms of local economic effects, we find limited evidence that these subsidized firms have employment spillovers in

[16] Kline and Moretti (2014b) find that the elasticity of local productivity with respect to population density is constant, suggesting that at least in the context of the Tennessee Valley Authority, the gains in one location are roughly offset by losses in the other. Some recent work on firm incentives in the United Kingdom finds effects on employment and investment, but little gain in total factor productivity (Criscuolo et al. 2019).

the local economy. In that case, the argument for this place-based policy rests more heavily on equity considerations.

Many questions remain unanswered. How much do these policies improve the well-being of underemployed and low-income workers? Are the most distressed places able to attract firms with tax incentives? How effective are these approaches relative to other policies? Does targeting subsidies at the largest firms have anticompetitive effects in the product market? At the local level, is the newly attracted firm stimulating hiring of local residents who were previously unemployed and working in low-wage jobs? Or as was argued in the case of Amazon's proposal for putting a headquarters in New York City, are all the good jobs going to people moving in from other locations, while leaving locals with more congestion and higher prices?

Policymakers can design incentives with these considerations in mind and evaluate the extent to which these policies actually "trickle down." Bartik (2019a) calls for targeting tax incentives to hard-hit regions and to employers who promise to hire local residents. He also notes that targeting marginal investments and job creation in tradable industries with high multipliers, instead of individual firms, could reduce political influence. More evidence on the conditions under which these policies are effective and for whom would help improve policy recommendations.

To the extent that well-targeted and effective policy is not feasible, recent harmonization efforts at the state and local level also hold promise. New York state lawmakers have proposed the End Corporate Welfare Act to outlaw firm-specific state tax incentives and have urged other states to do the same. Of course, an incentives truce is much more attractive to a state like New York than it is to states with distressed regions that are struggling to attract firms and grow their local economies.

Given disparate impacts of a universal ban on firm-specific incentives and regional heterogeneity, another promising avenue is harmonization at a regional level. Fajgelbaum et al. (2019) find that regional tax harmonization can achieve most of the gains of a national tax harmonization. In 2019, Kansas and Missouri came to a truce: they would not offer tax incentives for firms moving from the other side of the border in Kansas City. However, both states rushed to finalize large incentive deals right before the truce was enacted.

Avoiding these dynamics and directing funds to the most distressed regions may require a larger federal role. However, implementation details are key. Lenient eligibility requirements for the new federal opportunity zone program, for example, may not be as effective at targeting distressed regions as past programs with similar goals and stricter criteria like empowerment zones, which funded places with poverty rates above 40 percent and unemployment rates above 15 percent (Busso, Gregory, and Kline 2013b). Other countries have adopted more centralized approaches. For example, the European Union restricts state aid to reduce concerns about tax competition. Instead, they implement EU structural funds at the superfederal level and use incentives to reduce regional disparities by encouraging investment, capital deepening, and economic development in distressed areas.

Given the scale and scope of state and local business tax incentives in the United States, much more work needs to be done by academics and policymakers to analyze how these programs affect the welfare of local areas and the nation. While there have been recent efforts to increase transparency (Governmental Accounting Standards Board Statement 77), the new reporting requirements still give too much discretion to governments in terms of what to report and how to report it. The data are not yet uniform, comprehensive, or high quality. Even with new accounting rules, roughly half of municipalities have not disclosed any revenue lost to tax incentives in their annual financial reports (Farmer 2018). Meanwhile, at the state level, Michigan, Kansas, and Montana recently enacted laws requiring evaluation of business tax incentives, but many other states still do not have them.

■ *We are especially grateful for helpful comments on an early draft from Tim Bartik, Pat Kline, Ilyana Kuziemko, Juan Carlos Suárez Serrato, and Dan Wilson as well as the editorial team—Gordon Hanson, Enrico Moretti, Timothy Taylor, and Heidi Williams. We also thank Daron Acemoglu, Leah Boustan, Nick Buchholz, Janet Currie, Hank Farber, Thomas Fujiwara, Rick Hornbeck, Henrik Kleven, Carlianne Patrick, Jim Poterba, Steve Redding, Stefanie Stantcheva, Matt Turner, John Van Reenen, and Eric Zwick for helpful comments and discussions. Stephanie Kestelman and Dustin Swonder provided excellent research assistance. We thank Carlianne Patrick for providing us data on million-dollar plants. This work is supported by the National Science Foundation under grant 1752431. We declare that we have no relevant or material financial interests that relate to the research described in this paper.*

References

Acemoglu, Daron, David Autor, David Dorn, Gordon H. Hanson, and Brendan Price. 2016. "Import Competition and the Great US Employment Sag of the 2000s." *Journal of Labor Economics* 34 (S1): S141–98.

Agrawal, David, William Hoyt, and John D. Wilson. 2019. "Local Policy Choice: Theory and Empirics." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3545542.

Akcigit, Ufuk, John Grigsby, Tom Nicholas, and Stefanie Stantcheva. 2018. "Taxation and Innovation in the 20th Century." National Bureau of Economic Research Working Paper 24982.

Albouy, David. 2009. "The Unequal Geographic Burden of Federal Taxation." *Journal of Political Economy* 117 (4): 635–67.

Austin, Benjamin, Edward Glaeser, and Lawrence Summers. 2018. "Saving the Heartland: Place-Based Policies in 21st Century America." *Brookings Papers on Economic Activity* 48 (1): 151–255.

Baicker, Katherine, Jeffrey Clemens, and Monica Singhal. 2012. "The Rise of the States: US Fiscal Decentralization in the Postwar Period." *Journal of Public Economics* 96 (11–12): 1079–91.

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

Bartik, Timothy J. 2017. *A New Panel Database on Business Incentives for Economic Development Offered by State and Local Governments in the United States.* Kalamazoo: W.E. Upjohn Institute.

Bartik, Timothy J. 2018. "'But for Percentages for Economic Development Incentives: What Percentage Estimates Are Plausible Based on the Research Literature?" W.E. Upjohn Institute Working Paper 18-289.

Bartik, Timothy J. 2019a. *Making Sense of Incentives: Taming Business Incentives to Promote Prosperity.* Kalamazoo: W.E. Upjohn Institute.

Bartik, Timothy J. 2019b. "Should Place-Based Jobs Policies Be Used to Help Distressed Communities?" W.E. Upjohn Institute Working Paper 19-308.

Bennett, Tom. 2008. "Volkswagens Announcement of a New Tennessee Plant Can't Speed Work on Corridor K…" *Hiwassee River Watershed Coalition*, July 15. https://hrwc.net/volkswagens-announcement-of-a-new-tennessee-plant-cant-speed-work-on-corridor-k/.

Bilal, Adrien. 2019. "The Geography of Unemployment." https://drive.google.com/file/d/1DK9aTmJMD-DykuAoNUtGcdI4LfEWmhoK/view.

Bivens, Josh. 2019. *Updated Employment Multipliers for the U.S. Economy*. Washington, DC: Economic Policy Institute Report.

Black, Dan A., and William H. Hoyt. 1989. "Bidding for Firms." *American Economic Review* 79 (5): 1249–56.

Bloom, Nicholas, Erik Brynjolfsson, Lucia Foster, Ron Jarmin, Megha Patnaik, Itay Saporta-Eksten, and John Van Reenen. 2019. "What Drives Differences in Management Practices?" *American Economic Review* 109 (5): 1648–83.

Bruns, Adam. 2008. "Chattanooga Stands and Delivers: Volkswagen Chooses a Former U.S. Army Munitions Site for a Billion-Dollar Plant." *Site Selection Magazine*, September. https://siteselection.com/issues/2008/sep/Site-Visit/.

Bureau of Labor Statistics. 1990–2017a. "Local Area Unemployment Statistics." https://www.bls.gov/lau/ (accessed September 12, 2019).

Bureau of Labor Statistics. 1990–2017b. "Quarterly Census of Employment and Wages." https://www.bls.gov/cew/ (accessed March 7, 2019).

Busso, Matias, Jesse Gregory, and Patrick Kline. 2013a. "Assessing the Incidence and Efficiency of a Prominent Place Based Policy." *American Economic Review* 103 (2): 897–947.

Busso, Matias, Jesse Gregory, and Patrick Kline. 2013b. Institute for Research on Poverty, University of Wisconsin Magazine Newsletter/Report. https://www.irp.wisc.edu/publications/focus/pdfs/foc301d.pdf.

Chattanooga Times Free Press. 2008. "Chattanooga: VW Incentives Largest in State." July 24. tmcnet.com/usubmit/2008/07/24/3565003.htm.

Chirinko, Robert S., and Daniel J. Wilson. 2008. "State Investment Tax Incentives: A Zero-Sum Game?" *Journal of Public Economics* 92 (12): 2362–84.

Corasaniti, Nick, and Matthew Haag. 2019. "The Tax Break Was $260 Million. Benefit to the State Was Tiny: $155,520." *New York Times*, May 1. https://www.nytimes.com/2019/05/01/nyregion/nj-tax-break-kevin-sheehan.html.

Council of State Governments. 1950–2018. *Book of the States*. Lexington: Council of State Governments.

Criscuolo, Chiara, Ralf Martin, Henry G. Overman, and John Van Reenen. 2019. "Some Causal Effects of an Industrial Policy." *American Economic Review* 109 (1): 48–85.

Crisp, Elizabeth. 2016. "Lobbyists Spent Nearly Half a Million Dollars to Woo Louisiana Legislators in 2016." *Advocate*, December 18. https://www.theadvocate.com/baton_rouge/news/politics/legislature/article_e341e746-c3d8-11e6-aa43-7b3565cac1fb.html.

Curtis, E. Mark, and Ryan A. Decker. 2018. "Entrepreneurship and State Taxation." Finance and Economics Discussion Paper 2018–003.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger. 2019. "The Rise of Market Power and the Macroeconomic Implications." *The Quarterly Journal of Economics* (forthcoming). http://www.janeeckhout.com/wp-content/uploads/26.pdf.

Dharmapala, Dhammika. 2008. "What Problems and Opportunities Are Created by Tax Havens?" *Oxford Review of Economic Policy* 24 (4): 661–79.

Fajgelbaum, Pablo D., Eduardo Morales, Juan Carlos Suárez Serrato, and Owen Zidar. 2019. "State Taxes and Spatial Misallocation." *Review of Economic Studies* 86 (1): 333–76.

Farmer, Liz. 2018. "Despite New Rules to Disclose Corporate Tax Breaks, Just Half of Local Governments Are." *Governing*, March 21. https://www.governing.com/topics/finance/gov-new-rules-disclose-tax-breaks-half-governments.html.

Federal Housing Finance Agency. 2014. "FHFA HPI." https://www.fhfa.gov/DataTools/Downloads/Documents/HPI/HPI_AT_BDL_county.xlsx (accessed January 2, 2019).

Follow the Money. 2000–2016. https//www.followtheoney.org/.

Fuest, Clemens, Andreas Peichl, and Sebastian Siegloch. 2018. "Do Higher Corporate Taxes Reduce Wages? Micro Evidence from Germany." *American Economic Review* 108 (2): 393–418.

Garcia-Mila, Teresa, and Therese J. McGuire. 2002. "Tax Incentives and the City." *Brookings-Wharton Papers on Urban Affairs* 3: 95–132.

Garrett, Daniel G., Eric Ohrn, and Juan Carlos Suárez Serrato. Forthcoming. "Tax Policy and Local Labor Market Behavior." *American Economic Review: Insights*. https://doi.org/10.1257/aeri.20190041.

Gaubert, Cecile. 2018. "Firm Sorting and Agglomeration." *American Economic Review* 108 (11): 3117–53.

Gaubert, Cecile, Patrick Kline, and Danny Yagan. 2020. "Place-Based Redistribution." Unpublished.

Giroud, Xavier, and Joshua Rauh. 2019. "State Taxation and the Reallocation of Business Activity: Evidence from Establishment-Level Data." *Journal of Political Economy* 127 (3): 1262–1316.

Glaeser, Edward L. 2001. "The Economics of Location-Based Tax Incentives." Harvard Institute of Economic Research Discussion Paper 1932.

Glaeser, Edward L., and Joshua D. Gottlieb. 2008. "The Economics of Place-Making Policies." National Bureau of Economic Research Working Paper 14373.

Glaeser, Edward L., and Naomi Hausman. 2019. "The Spatial Mismatch between Innovation and Joblessness." National Bureau of Economic Research Working Paper 25913.

Glaeser, Edward L., Hedi D. Kallal, José A. Scheinkman, and Andrei Shleifer. 1992. "Growth in Cities." *Journal of Political Economy* 100 (6): 1126–52.

Glaeser, Edward L., José A. Scheinkman, and Andrei Shleifer. 1995. "Economic Growth in a Cross-Section of Cities." *Journal of Monetary Economics* 36 (1): 117–43.

Good Jobs First. 1976–2019. "Subsidy Tracker." https://www.goodjobsfirst.org/subsidy-tracker (accessed March 1, 2019).

Gordon, Roger H. 1983. "An Optimal Taxation Approach to Fiscal Federalism." *Quarterly Journal of Economics* 98 (4): 567–86.

Greenstone, Michael, and Enrico Moretti. 2003. "Bidding for Industrial Plants: Does Winning a 'Million Dollar Plant' Increase Welfare?" National Bureau of Economic Research Working Paper 9844.

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

Hall, Bronwyn, and Marta Wosinska. 1999. "The California R&D Tax Credit: Description, History, and Economic Analysis." A Report to the California Council on Science and Technology.

Haslam, Bill. 2011. *The Budget: Fiscal Year 2011–2012*. Nashville: Tennessee Department of Finance and Administration.

Helms, L. Jay. 1985. "The Effect of State and Local Taxes on Economic Growth: A Time Series–Cross Section Approach." *Review of Economics and Statistics* 67 (4): 574–82.

Henderson, J. Vernon. 2003. "Marshall's Scale Economies." *Journal of Urban Economics* 53 (1): 1–28.

Hines, James R. Jr. 1996. "Altered States: Taxes and the Location of Foreign Direct Investment in America." *American Economic Review* 86 (5): 1076–94.

Hirsch, Barry, David A. Macpherson, and Wayne G. Vroman. 1964–2018. "Union Density Estimates by State, 1964–2018." http://unionstats.gsu.edu/MonthlyLaborReviewArticle.htm (accessed July 29, 2019).

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

Jensen, Nathan M., and Edmund J. Malesky. 2018. *Incentives to Pander: How Politicians Use Corporate Welfare for Political Gain*. Cambridge University Press.

Karlin, Sam. 2018. "Chemical Sector Supports Nearly $80 Billion in Sales in Louisiana, Industry Study Says." *Advocate*, May 9. https://www.theadvocate.com/baton_rouge/news/business/article_e348431c-5396-11e8-aa28-f3bbd1d6f2a0.html.

Keen, Michael. 2001. "Preferential Regimes Can Make Tax Competition Less Harmful." *National Tax Journal* 54 (4): 757–62.

Kline, Patrick, and Enrico Moretti. 2014a. "Local Economic Development, Agglomeration Economies, and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority." *Quarterly Journal of Economics* 129 (1): 275–331.

Kline, Patrick, and Enrico Moretti. 2014b. "People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs." *Annual Review of Economics* 6 (1): 629–62.

Leachman, Michael. 2017. "Timeline: 5 Years of Kansas' Tax Cut Disaster." *Center on Budget and Policy Priorities*, May 24. https://www.cbpp.org/blog/timeline-5-years-of-kansas-tax-cut-disaster.

Markell, Jack. 2017. "Let's Stop Government Giveaways to Corporations." *New York Times*, September 21. https://www.nytimes.com/2017/09/21/opinion/incentives-businesses-corporations-giveaways.html.

Mast, Evan. Forthcoming. "Race to the Bottom? Local Tax Break Competition and Business Location."

American Economic Journal: Applied Economics. https://doi.org/10.1257/app.20170511.

Moretti, Enrico. 2010. "Local Multipliers." *American Economic Review: Papers & Proceedings* 100: 1–7.

Moretti, Enrico. 2019. "The Effect of High-Tech Clusters on the Productivity of Top Inventors." National Bureau of Economic Research Working Paper 26720.

Moretti, Enrico, and Daniel J. Wilson. 2014. "State Incentives for Innovation, Star Scientists and Jobs: Evidence from Biotech." *Journal of Urban Economics* 79: 20–38.

Moretti, Enrico, and Dan Wilson. 2017. "The Effect of State Taxes on Geographical Location of Top Earners: Evidence from Star Scientists." *American Economic Review* 107(7): 1858–1903.

Oates, Wallace E. 1972. *Fiscal Federalism.* New York: Harcourt Brace Jovanovich.

Ohrn, Eric. 2018. "The Effect of corporate taxation on investment and financial policy: evidence from the DPAD." *American Economic Journal: Economic Policy* 10(2): 272–301.

Ossa, Ralph. 2015. "A Quantitative Analysis of Subsidy Competition in the US." National Bureau of Economic Research Working Paper 20975.

Patrick, Carlianne. 2016. "Identifying the Local Economic Development Effects of Million Dollar Facilities." *Economic Inquiry* 54 (4): 1737–62.

Patrick, Carlianne, and Mark Partridge. 2019. "Identifying Agglomeration Spillovers: New Evidence from Large Plant Openings." Unpublished. https://www.dropbox.com/sh/w9ilyyom5te4hgf/AAB1n9Hz0wXb7etHdRHQ8f2Ba?dl=0.

Ramsey, Frank P. 1927. "A Contribution to the Theory of Taxation." *Economic Journal* 37 (145): 47–61.

Romer, Christina D., and David H. Romer. 2010. "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks." *American Economic Review* 100 (3): 763–801.

Slattery, Cailin Ryan. 2019. "Bidding for Firms: Subsidy Competition in the U.S." http://cailin-slattery-zt3j.squarespace.com/s/Slattery-2020-Bidding-for-Firms.pdf.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick. 2019. "Capitalists in the Twenty-first Century." *The Quarterly Journal of Economics* 134(4): 1675–1745.

Sollaci, Alexandre B. 2019. "Agglomeration, Innovation, and Spatial Reallocation: The Aggregate Effects of R&D Tax Credits." https://www.dropbox.com/s/3stcd7dp81cqbmq/JMP_ASollaci.pdf?dl=0.

Spathelf, Christof. 2011. "First Person: The Site Selection Process behind VW's First U.S. Manufacturing Facility." *AreaDevelopment*, November. https://www.areadevelopment.com/Automotive/November2011/VW-Christof-Spathelf-Overseas-manufacturing-77780192.shtml.

Storey, Louise, Tiff Fehr, and Derek Watkins. 2012. "United States of Subsidies." *New York Times*, December. https://archive.nytimes.com/screenshots/www.nytimes.com/interactive/2012/12/01/us/government-incentives.jpg.

Suárez Serrato, Juan Carlos, and Owen Zidar. 2016. "Who Benefits from State Corporate Tax Cuts? A Local Labor Markets Approach with Heterogeneous Firms." *American Economic Review* 106 (9): 2582–2624.

Suárez Serrato, Juan Carlos, and Owen Zidar. 2018. "The Structure of State Corporate Taxation and Its Impact on State Tax Revenues and Economic Activity." *Journal of Public Economics* 167: 158–76.

Summers, Lawrence H. 2019. "We No Longer Share a Common Lived Experience." *Washington Post*, October 9. https://www.washingtonpost.com/opinions/we-no-longer-share-a-common-lived-experience/2019/10/08/f037b9e4-e9eb-11e9-9c6d-436a0df4f31d_story.html.

Urban-Brookings Tax Policy Center. 1977–2017. "US Survey of State and Local Government Finances." https://slfdqs.taxpolicycenter.org/pages.cfm (accessed April 21, 2018).

US Bureau of Economic Analysis. 1967–2017. "Gross Domestic Product (GDP) by State." https://www.bea.gov/data/gdp/gdp-state (accessed January 21, 2019).

US Census Bureau. 1997–2017. County Business Patterns. https://www.census.gov/programs-surveys/cbp/data.html (accessed January 13, 2019).

US Census Bureau. 2000. "2000 U.S. Census."

Wilson, Daniel J. 2009. "Beggar thy neighbor? The in-state, out-of-state, and aggregate effects of R&D tax credits." *The Review of Economics and Statistics* 91(2): 431–436.

Journal of Economic Perspectives—Volume 34, Number 2—Spring 2020—Pages 119–142

Taxation and Migration: Evidence and Policy Implications

Henrik Kleven, Camille Landais, Mathilde Muñoz, and Stefanie Stantcheva

Tax rates differ substantially across countries and across locations within countries. An important question is whether people choose locations in response to these tax differentials, thus reducing the ability of local and national governments to redistribute income and provide public goods. Due to globalization and the lowering of mobility costs, it has become increasingly important to pay attention to mobility responses when designing tax policy. In this paper, we review what we know about mobility responses to personal taxation and discuss the policy implications. Our main focus is on the mobility of people, especially high-income people, but we will also discuss the mobility of wealth in response to personal taxes.

It is clear that high-income individuals sometimes move across borders to avoid taxes. The media is filled with examples of famous people who, often by their own admission, engage in such tax avoidance behavior. The Rolling Stones left England for France in the early 1970s in order to avoid the exceptionally high top marginal

■ *Henrik Kleven is Professor of Economics and Public Affairs, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, New Jersey. Camille Landais is Professor of Economics, London School of Economics, London, United Kingdom. Mathilde Muñoz is a PhD student, Paris School of Economics, Paris, France. Stefanie Stantcheva is Professor of Economics, Harvard University, Cambridge, Massachusetts. Kleven and Stantcheva are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are kleven@princeton.edu, c.landais@lse.ac.uk, mathilde.munoz@psemail.eu, and sstantcheva@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.34.2.119.

tax rates—well above 90 percent—in the UK at the time.[1] Many other British rock stars moved to lower tax jurisdictions, including David Bowie (Switzerland), Ringo Starr (Monte Carlo), Cat Stevens (Brazil), Rod Stewart (United States), and Sting (Ireland). In more recent years, actor Gérard Depardieu moved to Belgium and eventually Russia in response to the 75 percent millionaire tax in France, while a vast number of sports stars in tennis, golf, and motor racing have taken residence in tax havens such as Monte Carlo, Switzerland, and Dubai.

While these anecdotes are suggestive, two caveats prevent us from drawing any broader conclusion from them. First, all of the examples are from the sports and entertainment industries. These industries may feature particularly high cross-border mobility, both because they involve little location-specific human capital and because workers tend to be less tied to specific firms. Second, some of the examples reflect location responses to extreme top tax rates. The key question—and the one with which we are preoccupied in this paper—is if income tax rates distort the location choices of broader segments of workers? And if they do, how large are the responses and what are the implications for policy? These questions are particularly pertinent due to the recent proposals in the United States and elsewhere to raise the taxation of income or wealth substantially at the top of the distribution.

Mobility of People

The idea that tax policy may affect the location decisions of individuals has a long tradition in economics. In fact, tax-induced mobility is a central mechanism in several strands of economic theory. In the local public finance literature, starting with the seminal contribution of Tiebout (1956), migration responses to local taxes and public goods are the fundamental force that governs the sorting of individuals across jurisdictions. Since the contributions of Rosen (1979) and Roback (1982), the field of economic geography has focused on spatial equilibrium models in which the assumptions placed on migration elasticities are key determinants of the spatial allocation of factors and the geographic variation in prices. The optimal taxation literature has also emphasized that migration responses can have important effects on tax design and may trigger socially inefficient tax competition in uncoordinated tax settings (for example, Mirrlees 1982; Bhagwati and Wilson 1989).

Despite its importance in economic theory and its salience in the policy debate, empirical evidence on the responsiveness of individual location decisions to taxes has been remarkably scant. Table 1 provides a summary of the available literature, focusing on papers that estimate mobility responses to personal income taxes.

[1] Up until 1978, the United Kingdom imposed a top marginal tax rate on labor income equal to 83 percent and a top marginal tax rate on capital income that was even higher, a stunning 98 percent. Very few people had sufficiently high incomes to face these tax rates, but rock stars were among them. The 1966 song “Taxman” by The Beatles was an attack on these high tax rates: “There’s one for you, nineteen for me/Cause I’m the taxman, yeah I’m the taxman/Should five percent appear too small/Be thankful I don’t take it all/Cause I’m the taxman, yeah I’m the taxman.”

Table 1
Summary of Empirical Literature on Migration Responses to Personal Income Taxes

Citation	*Countries*	*Population*	*Tax variation*	*Main result*	*Preferred mobility elasticity*
A: International mobility					
Akcigit, Baslandze, and Stantcheva (2016)	8 OECD countries	Top 1% of inventors	Variation across/ within states over time	Top foreign inventors are strongly mobile internationally	Foreigners = 1 Domestics = 0.03
Kleven, Landais, and Saez (2013)	14 European countries	Top football players	Variation across/ within countries over time	Top foreign footballers are strongly mobile internationally	Foreigners = 1 Domestics = 0.15
Kleven et al. (2014)	Denmark	Immigrants in the top 1%	Variation by earnings within country over time	Top foreign earners are strongly mobile in Denmark	Foreigners = 1.6 Domestics = 0.02
B: Within-country mobility					
Agrawal and Foremny (2019)	Spain	Top 1% of population	Variation across Spanish regions over time	Top taxpayers are strongly mobile within Spain	0.85
Akcigit et al. (2018)	8 US states	All inventors	Variation across/ within states over time	Inventors strongly mobile within the US	Out-of-state = 1.23 In-state = 0.11
Feldstein and Wrobel (1998)	USA	Sample of full-time workers	Variation across US states	Wage changes fully offset tax changes across US states	∞
Liebig, Puhani, and Sousa-Poza (2007)	Switzerland	Population aged 21–64	Variation across Swiss municipalities over time	College graduates and foreigners are mobile within Switzerland	NA
Martinez (2017)	Switzerland	Top 1% in canton of Obwalden	Variation across Swiss cantons over time	Rich taxpayers are strongly mobile within Switzerland	2.0
Moretti and Wilson (2017)	USA	Top 5% of inventors	Variation across US states over time	Top inventors are strongly mobile across US states	1.8
Schmidheiny (2006)	Switzerland	Households in and around Basel	Variation across Swiss municipalities	Rich households more likely to move to low-tax municipalities	NA
Schmidheiny and Slotwinski (2018)	Switzerland	Foreigners below earnings threshold	Variation from duration threshold in tax scheme	Top earners are strongly mobile within Switzerland	NA
Young et al. (2016)	USA	Millionaires	Variation across US states	Millionaires only moderately mobile within the US	0.1

Interestingly, only a dozen papers or so provide direct evidence on such responses and most of these papers are very recent. Two empirical challenges can explain the paucity of empirical research in this area: the lack of suitable data on migration and the lack of credible tax variation for identifying causal effects.

Information on migration patterns combined with precise measures of earnings and tax rates in different locations is hard to come by. Traditional surveys either lack this type of information or are statistically underpowered due to small sample sizes. One way of circumventing this data limitation is to focus on alternative outcomes, such as wages, and test structural predictions of migration models under different assumptions about mobility. Feldstein and Wrobel (1998) provide an early example of this approach. Their premise is the following. In the absence of heterogeneity in preferences for different locations, a long-run equilibrium equalizes utility across locations for all individuals and therefore fixes the net-of-tax wage rate in each location. In this case, there is perfect mobility: an increase in the tax rate in a given location must be exactly offset by an increase in the wage, because otherwise every individual would move out of that location. Testing if the elasticity of wages with respect to the net-of-tax rate equals minus one is therefore a test of perfect mobility (that is, an infinite mobility elasticity). Using cross-sectional variation in the progressivity of state income taxes in the United States, Feldstein and Wrobel estimate very large wage responses to the net-of-tax rate and cannot reject an elasticity of minus one. However, their large standard errors imply that, in a number of specifications, they also cannot reject the opposite extreme of small or zero elasticities.

The recent literature has taken two different approaches to overcome these data challenges. The first approach is to focus on specific segments of the labor market for which detailed migration information is available from external sources. Examples include football (soccer) players where rich biographical information allows one to reconstruct migration patterns (Kleven, Landais, and Saez 2013), and inventors whose location decisions can be inferred from patent records (Akcigit, Baslandze, and Stantcheva 2016; Akcigit et al. 2018; Moretti and Wilson 2017). The second approach is to find contexts in which administrative data with information on migration is available. For example, researchers have used tax or social security records from countries with a federal structure where the internal migration across tax jurisdictions can be observed (Young et al. 2016; Martinez 2017; Agrawal and Foremny 2019). Another possibility is to study countries, typically Scandinavian countries, that keep migration records of all movements in and out of the country that can be linked to administrative tax records (Kleven et al. 2014).

Where suitable migration data is available, the next challenge relates to the tax variation used to estimate migration responses. This challenge is twofold. First, one needs to measure correctly the tax incentive that governs location decisions. As with other extensive-margin decisions, location decisions depend on the *average* rather than the marginal tax rate, and average tax rates are not always straightforward to calculate.[2] Moreover, for workers at the lower end of the income distribution, the relevant average tax rate depends, not just on the tax system, but also on the potentially complicated system of welfare and social insurance programs.[3] Second, one needs

[2] Estimating the elasticity of migration with respect to the net-of-tax rate relies on correctly measuring the change in the tax incentive (the denominator of the elasticity). Otherwise, the elasticity estimate will be biased, even if the reduced-form effect of the reform on migration (the numerator) is correctly identified.

[3] Despite a long-standing debate about “welfare magnets” (for example, Borjas 1999), there is very little conclusive evidence on mobility responses to welfare benefits by low-income people. Agersnap, Jensen, and Kleven (2019) provide some of the first causal evidence on welfare magnet effects using variation from a special immigrant welfare scheme in Denmark.

Table 2
Summary of Preferential Tax Schemes to Foreigners

Country	*Year of implementation*	*Income eligibility criterion*	*Duration of scheme*	*Preferential tax treatment*
Denmark	1991	Yes	3 years originally, now extended to 7 years	Flat income tax of 30% originally, now 27%
Finland	1999	Yes	2 years	Flat income tax of 35%
France	2004	No	5 years originally, now extended to 8 years	30% of taxable income is tax exempt
Italy	2011	No	5 years	70% of taxable income was exempt originally, now 50%
Netherlands	1985	Yes	5 years originally, now extended to 10 years	35% of taxable income was exempt originally, now 30%
Portugal	2009	No	10 years	Flat income tax of 20%
Spain	2005	Yes since 2010	6 years	Flat income tax of 24%
Sweden	2001	Yes	3 years	25% of taxable income is tax exempt

Note: In the Netherlands, the 35 percent ruling has been officially implemented by law in 1985, but was used in a nonformal way since the 1960s, and was based on a nonpublic internal resolution of the Dutch Revenue Service.

to find tax variation that is plausibly orthogonal to other factors affecting individual location choices—including local labor market conditions, local amenities, and public goods—and sufficiently large to generate effects that can be detected in the data.

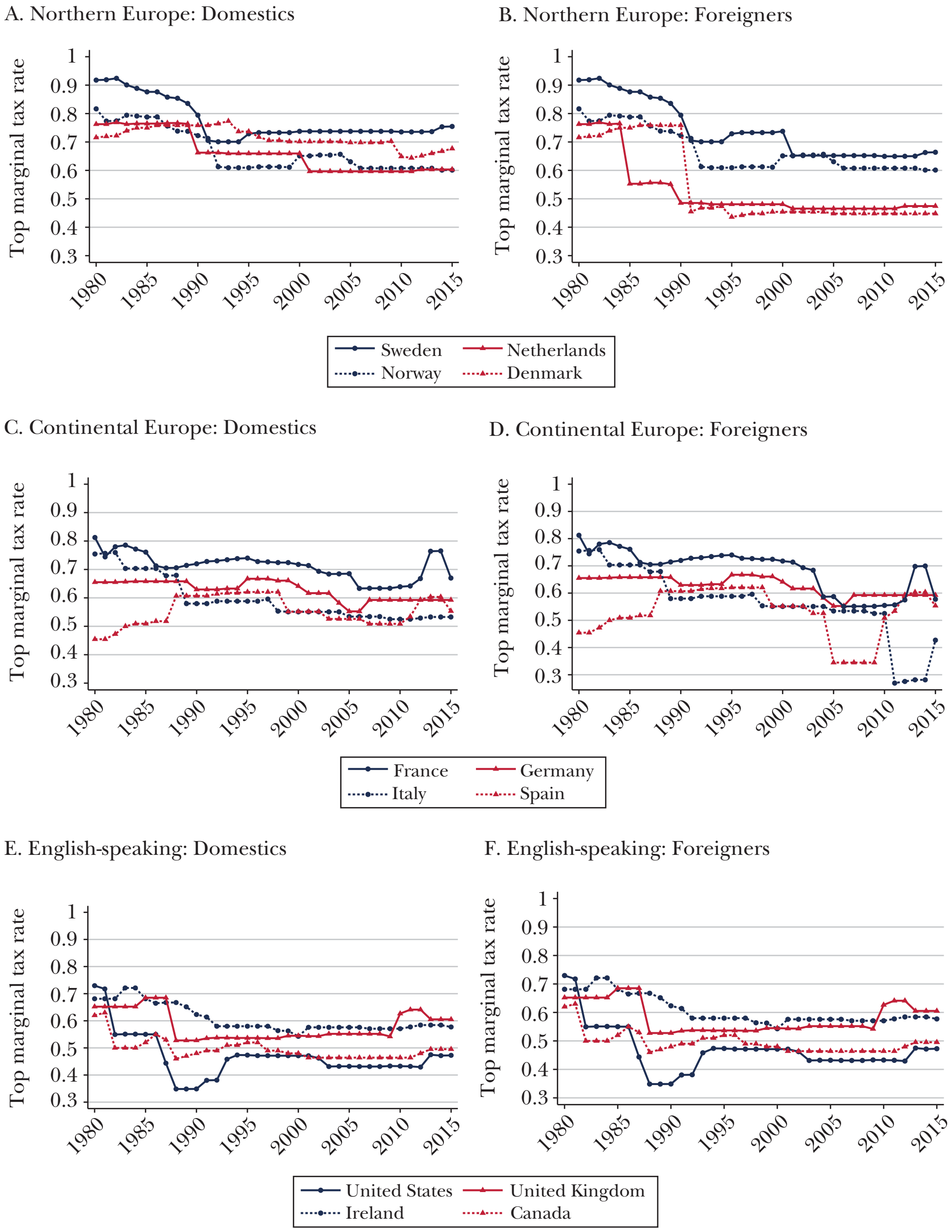
Motivated by these challenges, much of the recent literature has focused on people at the top of the earnings distribution. Beyond providing estimates of mobility responses for a segment of the population that may be particularly important for government revenue and economic efficiency, focusing on top earners offers important advantages.

First, for workers with very high earnings, the top marginal tax rate is a reasonable proxy for the average tax rate and is relatively easy to compute across countries and over time (Kleven, Landais, and Saez 2013). Specifically, the top marginal tax rate reflects the combined wedge from the top-bracket personal income tax rate, uncapped social security taxes on workers and firms, and consumption taxes (value-added, sales and excise taxes). Second, because of income tax reforms, top marginal tax rates exhibit substantial variation over time, both within and across countries, offering opportunities to identify the causal effect of taxes on migration. In particular, the introduction of preferential tax schemes to high-income foreigners in a number of countries provides useful sources of quasi-experimental variation for studying mobility responses (Kleven, Landais, and Saez 2013; Kleven et al. 2014). Table 2 provides details of existing preferential tax schemes in different countries, showing that they often introduce very large tax cuts for foreign over domestic workers.

To illustrate the identification opportunities created by the substantial variation in top marginal tax rates, Figure 1 plots their evolution in twelve countries

Figure 1

Top Marginal Tax Rates on Earnings 1980–2015



A. Northern Europe: Domestics

B. Northern Europe: Foreigners

C. Continental Europe: Domestics

D. Continental Europe: Foreigners

E. English-speaking: Domestics

F. English-speaking: Foreigners

Note: The figure depicts the evolution of top marginal tax rates on earnings in 12 OECD countries from 1980 to 2015. Our measure of top marginal tax rates includes top income tax rates, uncapped employer and employee payroll taxes, and consumption taxes. Top marginal tax rates on foreigners also account for the provisions of foreigners' tax schemes summarized in Table 2. See online Appendix A for details.

from 1980 to 2015.[4] The rows separate different sets of countries, while the columns distinguish between domestic and foreign residents. Several points are worth highlighting. First, the top marginal tax rate on domestic residents during this time period tends to be largest in northern Europe, intermediate in continental Europe, and smallest in English-speaking countries. For example, the top marginal tax rate equals 75 percent in Sweden and 48 percent in the United States in 2015. Second, this cross-country pattern is less pronounced when focusing on the taxation of foreigners, because preferential foreigner tax schemes are more prevalent in high-tax countries. Third, the introduction of preferential tax schemes to foreigners, due to their generosity and design, creates sharp variation in location incentives.

With this data in hand, a useful preliminary exercise consists in correlating the level of top marginal tax rates with the stock of migrants across countries. Beyond its descriptive purpose, this serves to illustrate the nature of the identification challenges and will also put the overall effect of taxes into perspective. Building on Muñoz (2019), we use survey data from the European Labour Force Survey (EU-LFS) and the Current Population Survey (CPS) to construct yearly measures of the stock of foreigners in 25 European countries and the United States between 2009–2015.[5] Because our focus is on high-income people, we select individuals with labor earnings in the top 5 percent of the distribution in each country and year. We then compute the log share of top 5 percent foreigners in the overall population, where foreigners are defined as citizens of a country different from their country of residence. In Panel A of Figure 2, we plot the average share of top 5 percent foreigners between 2009–15 against the average top marginal net-of-tax rate over the same period (both variables are measured in logs). The figure first confirms the large dispersion in tax rates across countries. On the far right of the diagram, eastern European countries like Bulgaria and the Czech Republic have high net-of-tax rates due to their flat income taxes with

[4] We combine the top personal income tax rate τ_i, the uncapped payroll tax rates on employees (workers) and employers (firms) τ_{pw} and τ_{pf}, and the VAT (or sales tax) rates τ_c in order to obtain our final measure of the top marginal tax rate τ:

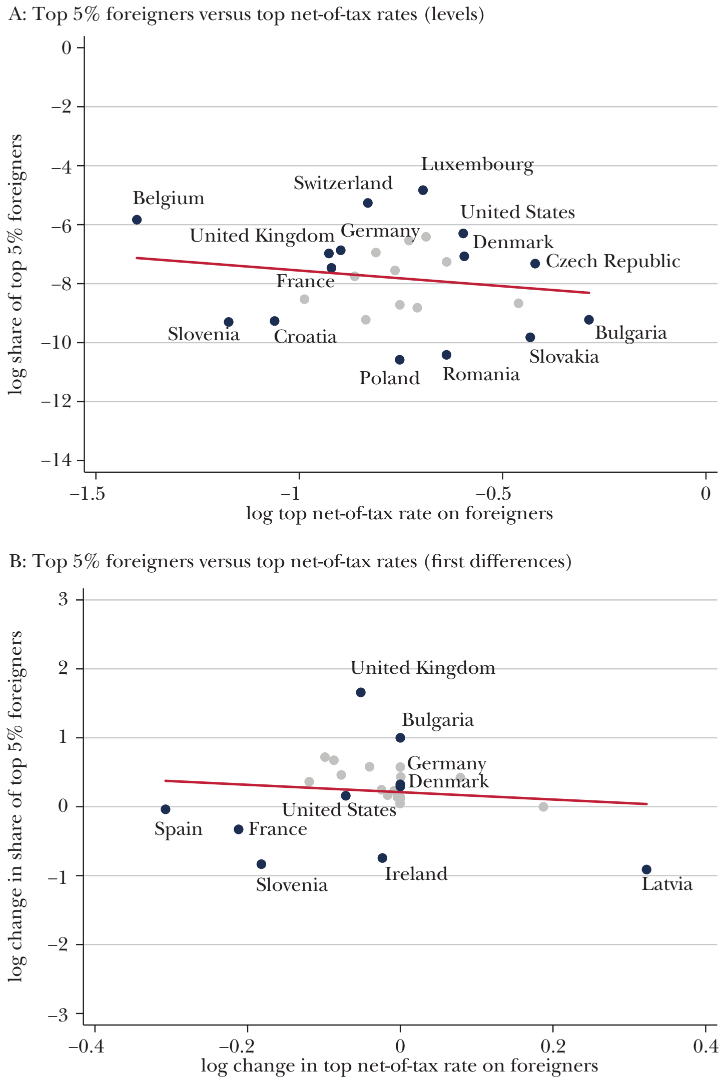
$$\left(1-\tau\right) = \frac{\left(1-\tau_i\right)\left(1-\tau_{pw}\right)}{\left(1-\tau_c\right)\left(1-\tau_{pf}\right)}.$$

Note that this formula has been written for the standard case where the employer's and employee's payroll taxes are both based on gross earnings and where the income tax rate applies to earnings net of all payroll taxes. When this is not the case, we have adapted our computations to capture precisely country-specific rules.

[5] The EU Labour Force Survey dataset is the largest European survey of individuals. It is a repeated cross section covering roughly 0.3 percent of the overall European population per year since the 1980s. It includes detailed income information since 2009. The full list of countries in our analysis is the following: Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Switzerland, the United Kingdom, and the United States. For top marginal tax rates, we extend the series created by Kleven, Landais, and Saez (2013) and Piketty, Saez, and Stantcheva (2014) to these 26 countries. We provide a complete description of the data and the construction of our sample in the online Appendix.

Figure 2

Cross-Country Evidence on Mobility Responses at the Top



A: Top 5% foreigners versus top net-of-tax rates (levels)

B: Top 5% foreigners versus top net-of-tax rates (first differences)

Note: The figure shows cross-country correlations between log shares of top earning foreigners and log top marginal net-of-tax rate on earnings for 25 European countries plus the United States. Shares of top earning foreigners are computed from the EU-Labor Force Survey and the Current Population Survey for the United States and defined as the number of foreigners who have earnings in the top 5 percent of the distribution divided by the total population of residents. The top marginal net-of-tax rate on earnings accounts for personal income taxes, uncapped payroll contributions, and consumption taxes. Panel A plots the average log share of top foreigners over the period 2009 to 2015 against the average log top marginal net-of-tax rate on earnings for foreign residents over the same period. Panel B plots the same correlation but in first-difference, focusing on variation between 2009 and 2015. Foreigners in the top 5 percent of the wage distribution represent 0.3 percent of the overall population in Belgium, 0.9 percent in Luxembourg, and 0.01 percent in Bulgaria. The corresponding level of top marginal tax rate is 75 percent for Belgium, 50 percent for Luxembourg, and 25 percent for Bulgaria for the same period. See text and online Appendix for details.

low rates. Interestingly, a country such as Denmark is also located on the far right of the diagram because of their preferential tax scheme to foreigners. Second, there is also a large dispersion in the share of foreign workers at the top of the earnings distribution. While countries like Luxembourg and Switzerland have large fractions of foreigners, eastern European countries have small shares. Most importantly, there is no sign of a positive correlation between the stock of foreigners and the net-of-tax rate. If anything, the correlation is negative: countries with large shares of foreigners at the top tend to be those with large tax rates at the top.

This figure lays bare that many country-specific factors affect migration decisions, and such factors must help to explain why countries such as Luxembourg, Belgium, or the United States attract a larger share of high-skill foreigners than Romania or Poland, despite having higher top tax rates on earnings. Furthermore, the factors that make a country attractive evolve significantly over time, as shown in panel B. There we move from a correlation in levels to a correlation in *changes* over time and ask if the share of top foreigners increases more (or falls by less) in countries that have reduced their top tax rate by more between 2009 and 2015. We find no correlation and see very different trends in the stock of top foreigners across countries with no variation in tax incentives. The United Kingdom, for instance, saw a large increase in the stock of top foreigners while Ireland experienced a significant decline—even though top tax rates were roughly constant in both countries.

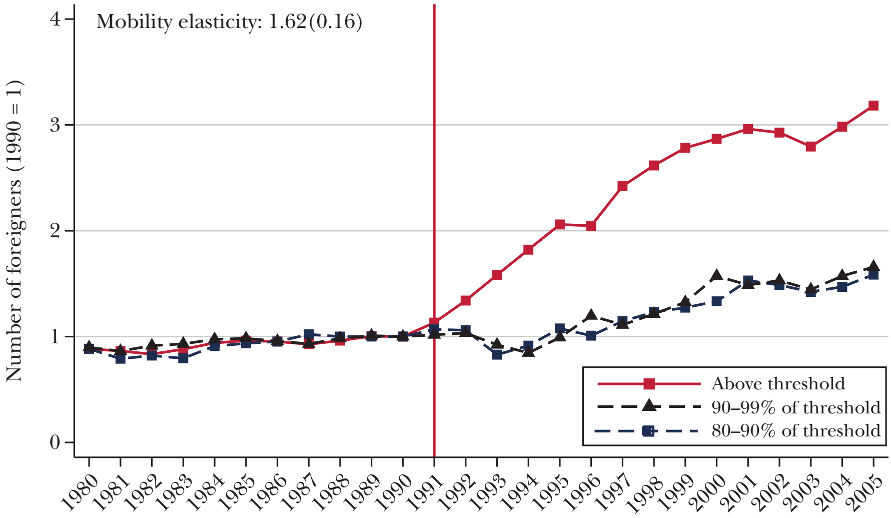
Controlling for all of the nontax determinants of location decisions that vary both across countries and over time is critical: any effect of taxes remains dominated and fully masked by such factors in the cross section.

Quasi-experimental Approaches

Quasi-experimental approaches leveraging variation in tax incentives across individuals within the same country over time can effectively control for any unobserved nontax determinants of location choices. Kleven, Landais, and Saez (2013) and Kleven et al. (2014) argue that the introduction of special tax schemes to foreigners provides such compelling quasi-experimental settings. Consider for instance the Danish tax scheme for foreigners, analyzed in detail by Kleven et al. (2014). This scheme was enacted in 1992 and applied to the earnings of foreign workers from June 1991 onwards. Eligibility for the scheme requires annual earnings above a threshold located around the ninety-ninth percentile of the earnings distribution. Initially, the scheme offered a flat income tax rate of 30 percent in lieu of the regular progressive income tax with a top marginal tax rate of 68 percent. The scheme could be used for a total period of up to three years, after which the taxpayer becomes subject to the ordinary income tax schedule. As shown in Table 2, the scheme parameters—tax rate, duration, and so on—have been revised since its introduction.

The design of the scheme lends itself to a difference-in-differences approach in which we compare the evolution of the number of foreigners above the eligibility threshold (treatments) and below the eligibility threshold (controls). Such an analysis is presented in Figure 3. It shows the stock of foreigners between 1980–2005 in the treated earnings range and in two untreated earnings ranges, between 80–90 percent

Figure 3
Migration Effects of the Danish Tax Scheme



Note: Originally produced by Kleven et al. (2014). The 1992 Danish tax reform, represented by a red vertical line, introduced a preferential tax scheme for foreign workers with earnings above an eligibility threshold, arriving in Denmark in or after 1991. The figure reports the evolution of the number of foreigners with earnings above the eligibility threshold from 1980 to 2005. It also reports the evolution of the number of foreigners in two control groups: individuals with earnings between 80 and 90 percent of the threshold and those with earnings between 90 percent and 99 percent of the threshold. All series are normalized to one in 1990, and numbers are weighted by duration of stay during the year for part-year foreign residents.

of the threshold and between 90–99 percent of the threshold. The two control series are normalized to match the treatment series in the pre-reform year. The graph provides exceptionally compelling evidence of mobility responses. The treatment and control series are perfectly parallel in the ten years leading up to the reform and start diverging immediately after the reform. The gap between the series builds up gradually through the 1990s and then reaches a steady state.[6] The effects are very large: the treatment series more than doubles relative to the control series, producing an elasticity of the stock of foreigners with respect to the average net-of-tax rate equal to about 1.6.

While the Danish evidence is very striking, it is important to highlight that mobility elasticities—as with other extensive-margin elasticities—are not structural

[6] The similarity between the two control series rules out the main potential confounder, namely that foreigners above the threshold are displacing foreigners just below the threshold. In that case, the divergence between treatments and controls would not represent a net mobility effect, but a shift in the earnings level of foreign arrivals. However, such shifting would produce a dip in the number of foreigners just below the threshold relative to the number of foreigners further down. The completely parallel trends of the two different control groups, along with other tests provided in Kleven et al. (2014), are inconsistent with such a story.

parameters. As a result, the findings in Kleven et al. (2014) are not necessarily transportable to other segments of the labor market or to other countries. To see that mobility elasticities can vary greatly across labor market segments, it is useful to explore heterogeneity across industries in the responses to the Danish tax scheme. Motivated by the many examples of tax-induced mobility in music, film, and sports discussed in the introduction, Figure 4 splits the difference-in-differences analysis into sports and entertainment (panel A) and all other industries (panel B). The effects are starkly different across these sectors. While the number of foreigners increased by a factor of less than two in other industries, it increased by a factor of more than five in sports and entertainment. Much of the dramatic increase in the latter group was driven by sports and, in particular, football (soccer) as analyzed by Kleven, Landais, and Saez (2013).[7]

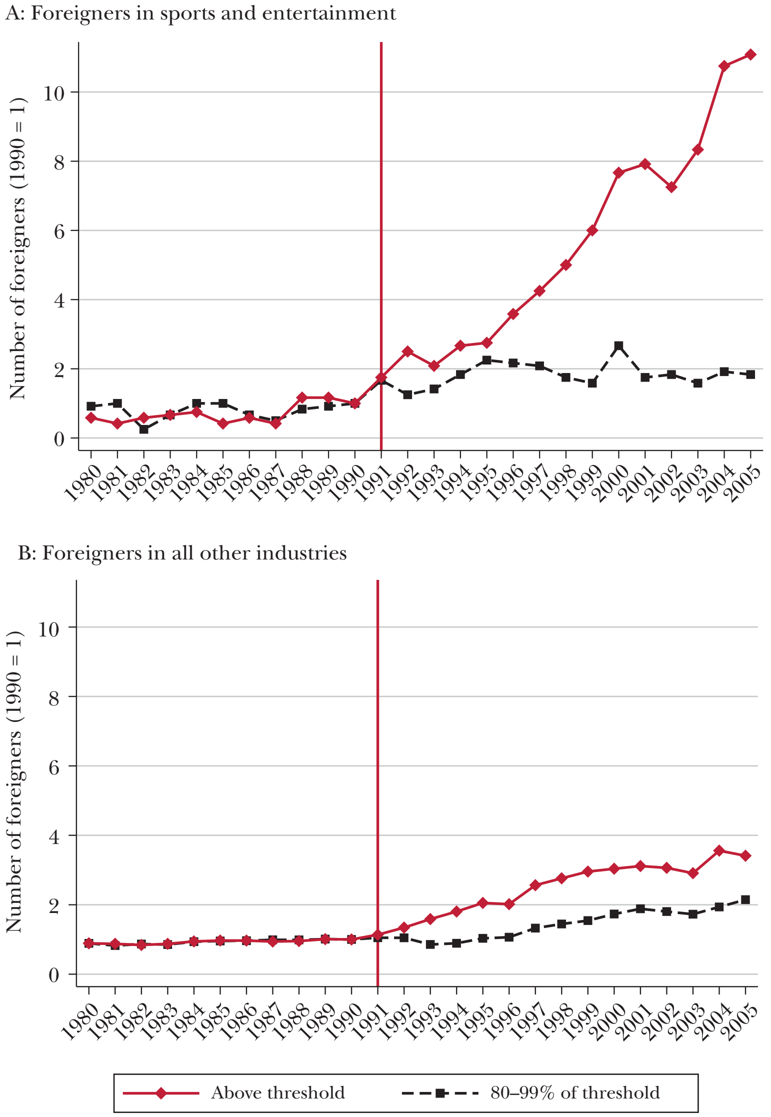
It is important to note that the mobility responses discussed above pertain to the immigration decisions of *foreign* citizens as opposed to the emigration or return-migration decisions of *domestic* citizens. The Danish scheme allows for studying the return-migration channel, because the scheme is available to any worker—independently of citizenship—who has been a tax resident abroad for at least three years (under the initial rules) prior to claiming the scheme treatment. As shown in Table 1, Kleven et al. (2014) find that the mobility elasticity of Danish expatriates is extremely small. Other papers that were able to identify the mobility elasticities of foreigners and domestics separately (Kleven, Landais, and Saez 2013; Akcigit, Baslandze, and Stantcheva 2016) also find much smaller elasticities for domestic workers. This difference can be explained, at least in part, by the fact that extensive-margin elasticities depend on the initial base. In any country, the vast majority of workers are domestic citizens rather than foreign citizens. As a result, the elasticity of foreign immigration represents a percentage change in an initially small stock of foreigners, whereas the elasticity of domestic emigration or return-migration represents a percentage change in an initially large stock of domestics. This mechanical difference in elasticities is very important for tax policy implications, as we discuss later.

Mobility elasticities are likely to vary not only by occupation or citizenship status, but also across countries within the same segment of the labor force. Indeed, an important question to ask is whether mobility elasticities are large only in small countries, for the same mechanical reasons that drive the differences in elasticities between foreigners and domestic residents. Can elasticities be sizable even for large countries that start with a large base of foreigners? Akcigit, Baslandze, and Stantcheva (2016) shed light on this question. They study the effects of top tax rates on the international mobility of "superstar" inventors—those with the most and best patents. Leveraging panel data from the US and European Patent Offices, they are able to track inventors over time and across countries and to exploit the differential impact

[7] The fact that immigration in the sports and entertainment industry starts accelerating after four years can be explained partly by regulation in the football sector until 1995. In addition, some sluggishness in the ability of firms (such as football clubs) to take full advantage of the scheme is natural due to information and hiring/firing frictions.

Figure 4

Migration Effects of the Danish Tax Scheme by Industry



Note: Originally produced by Kleven, Landais, and Saez (2013). The 1992 Danish tax reform, represented by a red vertical line, introduced a preferential tax scheme for foreign workers with earnings above an eligibility threshold, arriving in Denmark in or after 1991. The figure reports the evolution of the number of foreigners with earnings above the eligibility threshold separately for the sports and entertainment sector (panel A) and all other industries (panel B). In each panel, we also report the evolution of the number of foreigners in a control group of individuals with earnings between 80 and 99 percent of the threshold. All series are normalized to one in 1990, and numbers are weighted by duration of stay during the year for part-year foreign residents.

of top tax rates on inventors at different productivity and therefore income levels. They provide several country case studies, two of which are reproduced in Figure 5. Panel A considers once again the introduction of the Danish preferential tax scheme to foreigners, while panel B considers the US Tax Reform Act of 1986 which sharply reduced the top marginal income tax rate. Both panels rely on a synthetic control method, where a synthetic country is constructed as a weighted average of the other countries in the sample, in order to best fit the pre-reform time series of the treated country. The Danish reform shows a significant effect on the share of foreign inventors in Denmark, although the mobility elasticity is smaller than that estimated by Kleven et al. (2014) for the full population of high-income workers.[8] Importantly, the bottom panel suggests that the US Tax Reform Act of 1986 had a strong effect on the growth of foreign superstar inventors. In fact, the estimated mobility elasticity of top 1 percent superstar inventors for the US economy is extremely large, above 3.

In a complementary paper, Moretti and Wilson (2017) consider the mobility responses of star scientists across US states—rather than across countries—over the period 1977–2010. They estimate large long-run elasticities of mobility with respect to both personal and corporate income taxes. The elasticity of mobility with respect to personal income taxes is equal to 1.8.

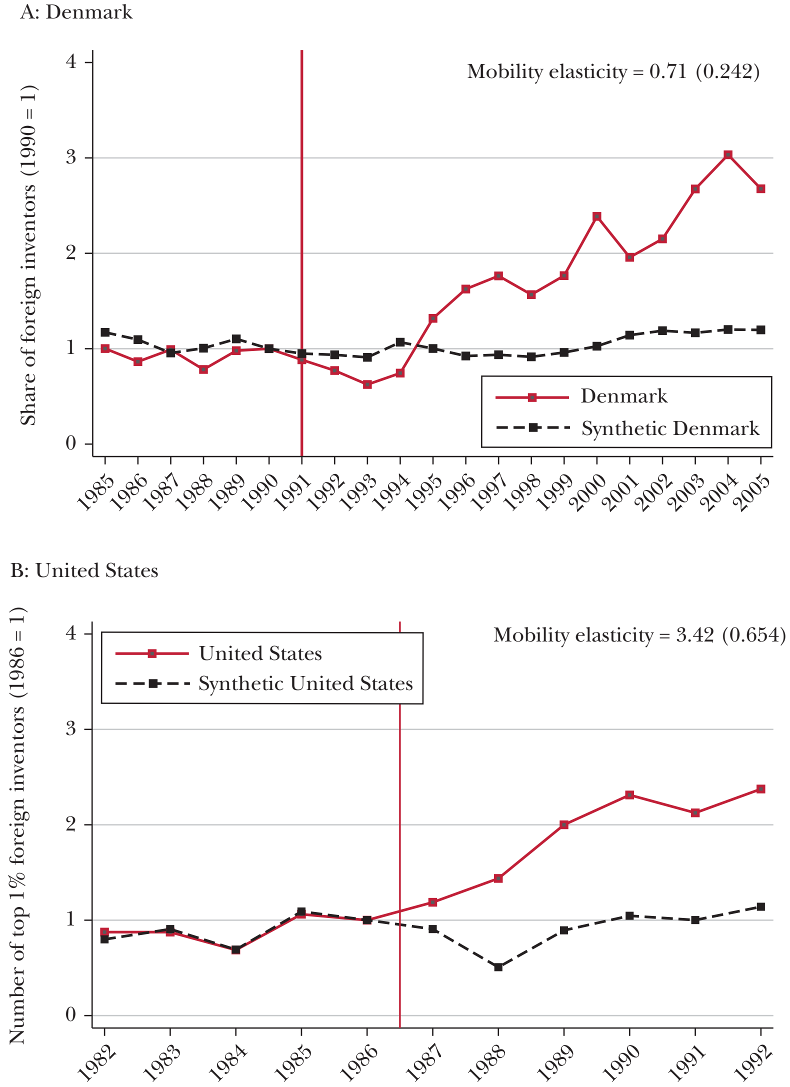
Are these tax-induced mobility effects only relevant for modern day economies? New historical evidence from Akcigit et al. (2018) shows significant effects of taxes on mobility across US states. They study the effects of personal and corporate taxes over the twentieth century in the United States, using a new panel of the universe of inventors who patented since 1920; a dataset of the employment, location, and patents of firms active in research and development since 1921; and a historical state-level corporate and personal tax database since 1900. They estimate that, over the twentieth century, the elasticity of the number of inventors residing in a state equals 0.11 for inventors from that state and 1.23 for inventors not from that state. These findings are consistent with the distinction made above, in the contest of international migration, between the mobility elasticities of foreigners and domestics.

Mobility of Wealth

So far, we have considered mobility responses to the taxation of labor income. However, mobility responses may depend on the tax treatment of both labor income and capital income, or on wealth. In general, it is easier to measure tax rates on labor income than on capital income and wealth. For the latter, detailed information on the income and wealth composition of individuals and their spouses is often

[8] Contrary to the effects on other occupations considered above, there is a lag in the effects of the reform on inventors. This can be explained by the fact that an inventor not only has to move to Denmark but also patent there, in order to be recorded as having moved to Denmark. Note also that the elasticity here can be diluted, because the analysis lumps together inventors across all income levels, some of which are not eligible for the foreign tax scheme (income is not observable in the patent data).

Figure 5
Migration Responses by Inventors



Note: Originally produced by Akcigit, Baslandze, and Stantcheva (2016). The figure shows inventors' migration response to two major tax reforms in Denmark and the United States. Panel A focuses on the 1992 Danish reform, which introduced a preferential tax scheme for foreign workers with earnings above an eligibility threshold, arriving in Denmark in or after 1991. The panel depicts the evolution of the share of foreign inventors, normalized to one in 1990, in Denmark and in a synthetic control country, constructed as a weighted average of all other countries in the sample, in order to best match the pre-reform series for Denmark. Panel B focuses on the 1986 Tax Reform Act, which lowered top marginal income tax rates in the United States. The panel shows the number of foreign inventors belonging to the top 1 percent of the distribution of citation-weighted patents in the United States and in a synthetic control country. Both series are normalized to one in 1986.

necessary. Absent such data, focusing on the top marginal tax rate on labor earnings can introduce measurement error in the true tax incentive, especially because some countries actively rely on specific provisions of capital taxation to attract foreign residents. Belgium is a case in point. Although its top marginal tax rate on earnings is relatively high, the fact that capital gains are exempt from taxation and inter vivos gifts are taxed at very low rates are often cited as reasons why a large number of wealthy French people have moved their tax residence to Uccle or Ixelles in the suburbs of Brussels.

Unfortunately, we still know little about the effects of capital income and wealth tax provisions on mobility. The data and identification challenges outlined above are, if anything, even stronger when studying the effects of wealth or estate taxation on migration. The literature is therefore limited to just a few papers. Most of them focus on within-country mobility responses to the taxation of bequests. In the United States, Bakija and Slemrod (2004) and Conway and Rork (2006) find that high inheritance and estate taxes have statistically significant, but modest, negative impacts on the number of federal estate tax returns filed in a state. This finding is echoed by Brülhart and Parchet (2014), who find that variation in bequest taxes across Swiss cantons have little impact on the location decisions of retirees. But the recent work by Moretti and Wilson (2019), focusing on individuals from the Forbes 400 in the United States, suggests that mobility responses to estate tax incentives might be larger at the very top of the wealth distribution.

An emerging literature also estimates taxable wealth elasticities, which in effect provides an upper bound on the size of the household mobility elasticity. Using a large wealth tax reform in Denmark, Jakobsen et al. (forthcoming) estimate long-run elasticities of taxable wealth with respect to the net-of-tax return at the top of the wealth distribution. They find sizable elasticities, between 0.7 and 1. Using variation in wealth taxation across Swiss cantons, Brülhart et al. (2019) also find large taxable wealth responses. They argue that these responses are not driven by the geographic mobility of people across cantons.

However, mobility responses to wealth and capital taxes are not limited to the mobility of people, because taxpayers may be able to relocate wealth and capital income without changing personal residence. In a perfectly enforced residence-based tax system, unless the individual owner changes fiscal residence, the geographic location of capital has no impact on tax liability. However, since residence-based taxation of capital income and wealth is difficult to enforce internationally, there is in practice considerable scope for tax avoidance through geographic mobility of capital income and wealth (for example, see Griffith, Hines, and Sørensen 2010). In general, we would expect capital to be more mobile than people, because it is less affected by the possibly strong and idiosyncratic preferences for specific locations.

The early empirical literature on international capital and wealth mobility relied predominantly on cross-country correlations. This body of work has been summarized by Gordon and Hines (2002) and Griffith et al. (2010). They argue that international tax provisions have significant effects on capital allocation, that tax avoidance through foreign investments and wealth holdings is a key threat to revenue collection and income redistribution, and that these forces have important implications for tax design.[9]

The most direct evidence on tax-related wealth mobility comes from recent work that uses creative data sources to quantify the amount of unreported wealth held in tax havens. Alstadsæter, Johannesen, and Zucman (2019) use leaked data from HSBC Switzerland and Mossack Fonseca (the so-called "Panama Papers") linked to administrative wealth records in Scandinavian countries. They show that the probability of hiding assets offshore rises sharply at the extreme tail of the wealth distribution: the top 0.01 percent of the wealth distribution owns about half of the leaked offshore wealth. Combining the micro data from specific leaks with estimates of the aggregate amount of offshore wealth from macroeconomic statistics (Zucman 2013), they argue that the top 0.01 percent evades about 25 percent of its tax liability by moving assets and investment income abroad. They do not estimate the response of such offshoring behavior to tax changes, but their evidence is certainly consistent with the presence of substantial tax-induced mobility of wealth by the very wealthy.

Policy Implications

The policy implications of mobility responses to taxes will depend on the extent to which tax policy is uncoordinated across countries—that is, a situation in which each country sets its tax rate without any international constraints or cooperation—and situations in which there is an element of coordination. We first consider uncoordinated tax policy and then turn to the implications of policy coordination.

Uncoordinated Tax Policy

When tax policy is not coordinated across countries, a key issue is whether there is targeted taxation of foreigners, as with the many preferential tax schemes discussed above, or a population-wide tax scheme applying to both foreign and domestic individuals. We start by discussing tax policy targeted to foreigners.

If the social welfare objective of a given country depends only on its domestic residents, the optimal influx of foreigners is governed solely by the externalities they generate on the domestic residents. As a result, the optimal taxation of foreigners trades off the revenue losses from cutting taxes on immigrants against the externality gains of attracting additional immigrants. These externalities include fiscal externalities—the additional tax revenue collected due to immigration—and non-fiscal externalities such as productivity spillovers (positive) and congestion (negative). In the absence of any non-fiscal externalities, the policy prescription is particularly

[9] Saez and Stantcheva (2018) derive the optimal tax rates on capital in different settings, including when capital income can be shifted abroad and there are different types of capital assets with potentially different elasticities.

simple: the optimal tax rate on foreigners is equal to the Laffer rate—that is, the tax rate that brings in the most revenue. In other words, if the government does not care about the welfare of the foreigners themselves and if the only externalities they create operate through the government budget, then it is optimal to maximize the revenue collected from them (Kleven, Landais, and Saez 2013).

These ideas can be formalized in a relatively straightforward manner. The theory of optimal taxation with migration responses was first analyzed by Mirrlees (1982) and reviewed by Piketty and Saez (2013) for the case without non-fiscal externalities, while the theory of welfare analysis with non-fiscal externalities was recently laid out in Kleven (2018). To simplify the analysis, let us make two assumptions. First, suppose the only behavioral response by foreigners is the migration response; labor supply conditional on moving is fixed. Second, suppose that the marginal non-fiscal externality from foreign immigration is zero. These assumptions are quite strong, but they provide a useful benchmark for developing intuition. Under these assumptions, we can show that the optimal tax rate on foreigners, τ^F, is given by

$$\tau^F \equiv \frac{1}{1+\eta^F},$$

where η^F is the elasticity of the stock of foreigners with respect to the net-of-tax rate.[10] The elasticity parameter η^F corresponds to the estimates reported in Table 1. This tax formula corresponds to the well-known inverse elasticity formula for the Laffer rate.

As mentioned, it is possible that foreign immigrants generate other externalities than those operating through the government budget. In particular, the foreigners targeted by the special tax schemes described above—high-income workers, researchers, and scientists—are often considered to have especially high positive spillovers. If such positive spillovers are taken into account, the optimal tax rate on foreigners would be correspondingly lower.

This discussion, along with evidence reviewed above, highlights the temptation of introducing preferential tax schemes for foreigners. For example, based on the tax scheme to foreigners in Denmark, Kleven et al. (2014) estimates a mobility elasticity of 1.6, which under the assumptions above implies a tax rate on foreigners of only 38 percent. While this is higher than the scheme's income tax rate, it is lower than the total top marginal tax rate when accounting for social security taxes

[10] This optimal tax rule in the equation can be derived as follows. Given the assumption of separability between the externalities from foreigners and the behavior of domestics, we define the external welfare effect of foreigners as $E^F(y^F N^F)$, where y^F denotes the average earnings of foreigners and N^F denotes the number of foreigners. The fact that we write the externality as a function of the aggregate earnings of foreigners, $Y^F \equiv y^F N^F$, as opposed to the number of foreigners is not crucial. Given foreigners are taxed separately at a flat tax rate of τ^F, the revenue collected from foreigners equals $\tau^F y^F N^F$. Denoting by μ the marginal value of government revenue, the government objective is to maximize $E^F(y^F N^F) + \mu \cdot \tau^F y^F N^F$. Absent intensive-margin responses (y^F is fixed), this yields the first-order condition for τ^F equal to $(\partial E^F / \partial Y^F) dN^F + \mu(d\tau^F N^F + \tau^F dN^F) = 0$. Defining the mobility elasticity as $\eta^F \equiv (dN^F/N^F)/(d(1-\tau^F)/(1-\tau^F))$ and the marginal externality effect in terms of the marginal value of government revenue as $e^F \equiv (\partial E^F/\partial Y^F)/\mu$, we would obtain the optimal tax rule shown here.

and value-added taxes (shown in Figure 1). Therefore, despite its apparent generosity, the Danish tax rate for foreigners may actually be higher than optimal. If we believe that top-earning foreigners coming to Denmark generate other positive externalities, then the optimal tax rate is even lower. In fact, the Danish tax scheme was originally motivated, not primarily by fiscal externalities and the Laffer logic, but by concerns about "brain drain" and the importance of high-skilled labor for economic growth and competitiveness. Our estimates imply that the fiscal externalities alone could justify Denmark's current preferential tax rate for foreigners.

While these arguments would seem to justify the use of preferential tax schemes to foreigners, a number of important qualifications should be emphasized. First, because mobility elasticities are not structural parameters, they may vary widely across countries and time periods. In particular, mobility elasticities depend mechanically on the size of the tax jurisdiction: a smaller jurisdiction is easier to exit. Indeed, as the size of a jurisdiction becomes infinitesimal, the mobility elasticity goes to infinity. Conversely, as the size of the jurisdiction approaches the global economy, the mobility elasticity goes to zero. Consistent with these conceptual ideas, the recent evidence showing large mobility responses is based predominantly on small tax jurisdictions: Denmark, Spanish regions, Swiss cantons, and US states. But the incentive to offer low taxes to foreigners is stronger in small countries such as Denmark than in large countries such as the United States.[11] By the same logic, the incentive to offer low taxes is stronger in subnational tax jurisdictions (a municipality or a state) than in a nation as a whole. The mechanical relationship between mobility elasticities and jurisdictional size can explain why tax havens tend to be small countries (Kanbur and Keen 1993).

Second, we have characterized the optimal tax policy of a given country not accounting for the welfare impact on other countries. Indeed, this formulation of the issues involves beggar-thy-neighbor policies done at the expense of other countries (although the externalities do not have to be symmetric, so the game is not exactly zero-sum). Moreover, in the case of special tax schemes targeted to foreign residents—unlike broader setting looking at taxes and provision of public goods—there is no clear Tiebout-sorting argument to justify the policy.

Third, the tax policy characterized above takes the policies of other countries as given. As analyzed in the literature on tax competition (for example, see Keen and Konrad 2013), when one country lowers its tax rate, other countries have an incentive to lower their tax rate too. But considering the tax rate series in Figure 1, there is no clear indication of a race to the bottom. Following an international trend of reducing top marginal tax rates around the 1980s, tax rates have remained relatively flat for the last two or three decades. Some countries have introduced special tax schemes to foreigners, but there is no evidence of any broad-scaled retaliation or race to the bottom. This might be because these preferential tax rates for foreigners

[11] A potentially offsetting effect is that negative congestion externalities—in the terms used above, the nonmonetary externalities from additional foreigners are negative rather than positive—are likely to be stronger in small countries.

have been introduced mostly in high-tax countries and are therefore perceived as leveling the playing field, rather than creating an unfair tax advantage. Still, it is interesting that almost all of the northern European countries have now introduced some version of a special tax scheme to foreigners, which suggests the possibility of tax competition between similar countries located in close proximity.

Finally, the policy implications change drastically if, instead of targeted taxation of foreigners, we consider uniform taxation of foreigners and domestic residents. Again, under the simplifying assumption that migration is the only behavioral response, the Laffer rate in an undifferentiated tax system equals $1/(1+\eta)$, where η is the average mobility elasticity on all residents. Because domestic residents constitute the vast majority of the population in most countries, η is approximately equal to the mobility elasticity of domestic residents. As shown in Table 1, this elasticity is very close to zero, and therefore, the Laffer rate is very close to one. Of course, there might be intensive-margin responses like reduced labor effort that lower the Laffer rate, but the key point here is that mobility responses across countries are not important for tax policy design unless the tax system targets foreign citizenship. This is not necessarily true of mobility responses across tax jurisdictions within countries like US states or Swiss cantons, where the relevant mobility elasticity may be considerably larger.

Coordinated Tax Policy

In the case of uncoordinated tax policy, each fiscal authority ignores any externalities that it imposes on other fiscal authorities (for example, Gordon 1983). A broadly coordinated tax policy is unlikely to materialize in the near future, even in otherwise integrated areas such as the European Union, both because fiscal policy is considered a matter of national sovereignty and because the potential gains from international tax coordination may be unevenly spread (Griffith, Hines, and Sørensen 2010). However, we can contemplate what such a policy would look like.

The issue of coordinated tax policy encompasses two main aspects. The first aspect concerns the level at which such coordination can happen. This leads to the question of the optimal size of jurisdiction over which tax policy is coordinated: for example, should it include a collection of countries (such as the European Union) or a collection of states within a country (such as the United States). The second aspect concerns what parts of fiscal policy are coordinated and to what degree.

On the first issue, a literature on fiscal federalism has studied the efficiency trade-offs associated with jurisdictional size (Oates 1972, 1999). Smaller jurisdictions (as mentioned above) will face larger migration elasticities and thus be more constrained in their choice of fiscal policy. They will have an incentive to lower tax rates, as in the earlier example of the special foreigner tax schemes. On the other hand, larger jurisdictions will be less able to cater to the diverse preferences for public goods and services among their residents. Diversity of policies, which may be valuable, could be lost. As a result, there is a trade-off between the inefficiencies from tax competition and the inefficiencies from public goods provision. Another challenge for large jurisdictions is an aversion to redistributing to immigrants in the

European Union and the United States, which can limit the ability to set progressive tax policy in a large and ethnically diverse jurisdiction (Alesina, Miano, and Stantcheva 2018). There may also be political economy frictions and transactions costs from administering large jurisdictions, which limit the ability of many countries to coordinate their tax policies.[12]

Regarding the type and degree of coordination, a conceptual distinction arises between situations where jurisdictions are constrained to set uniform policies and situations where they can—in a coordinated fashion—target taxes, transfers, and public goods to the local preferences of each jurisdiction. In the United States, for example, the federal government shoulders the bulk of progressive taxation, but states and municipalities have additional taxes, transfers, and public goods available to cater to their residents.

To formalize the conceptual ideas, consider a central tax authority such as a federal government or a supernational authority who sets tax policy in two regions, which we denote by A and B. To begin with, suppose the tax authority can set different tax rates in the two regions, τ^A and τ^B. We define two migration elasticities: η^A is the (positive) elasticity of migration in region A with respect to the net-of-tax rate in that region, while η^B is the (negative) elasticity of migration in region B to the net-of-tax rate in region A. Let's also assume that g^A is the average, income-weighted value to the social planner of transferring one unit of income to people in region A, while y^A and y^B denote aggregate incomes in the two regions. For simplification, assume that migration responses are the only behavioral responses to taxation, as we did in the previous section, and that any non-fiscal externalities are zero-sum across the two regions. Finally, assume that the aggregate tax revenue is rebated in a lump-sum fashion to all residents in the two regions, although this assumption can easily be relaxed. With this structure, it is possible to derive an optimal tax rate in region A:

$$\tau^A = \frac{1 - g^A - \tau^B \eta_A^B \cdot y^B / y^A}{1 - g^A + \eta^A}.$$

The formula for τ^B is symmetric.[13]

[12] There is also a small literature on the optimal size of countries more generally (Alesina and Spolaore 1997), which highlights the trade-offs between the efficiencies and inefficiencies from size. The trade-offs determining the optimal country size are between economies of scale from size (of which a reduced migration elasticity is a special case) and the gains from a diversity of policies adapted to residents' heterogeneous preferences.

[13] This formula is derived as follows. Conditional on moving to region A or B, person i has heterogeneous, but exogenously given income y_i^A or y_i^B. The total income in each region is then $y^A \equiv \sum_{i \in A} y_i^A$ and $y^B \equiv \sum_{i \in B} y_i^B$. As people can freely migrate, the income in each region is a function of both net-of-tax rates, i.e., $y^A = y^A(1 - \tau^A, 1 - \tau^B)$ and $y^B = y^B(1 - \tau^A, 1 - \tau^B)$. The central authority rebates the total tax revenues in a lump-sum fashion to all residents of the jurisdiction (this assumption can easily be relaxed). Thus, the consumption of agent i in region A under this tax system is $c_i^A = y_i^A(1 - \tau^A) + \tau^A y^A + \tau^B y^B$. People can have idiosyncratic preferences over the regions. Note, g_i is the marginal social welfare weight on agent i to be interpreted as a generalized social welfare weight as in Saez and Stantcheva (2016). Let

This formulation clarifies three main distinctions of coordinated policy relative to the uncoordinated policy setting considered in the preceding section. First, any non-fiscal externalities are internalized by the central tax authority, which no longer tries to implement beggar-thy-neighbor policies to benefit one region at the expense of the other. If these externalities are zero-sum, as assumed here, they drop out of the optimal tax formula entirely. Second, the central tax authority also internalizes the fiscal externalities that occur when people move between the two regions. This fiscal externality appears in the last term of the numerator and depends on the (negative) cross-elasticity of migration between the two regions and on the level of taxes in the other region. This term makes optimal taxes higher in both regions, all else being equal. Finally, the formula illustrates why it is valuable to differentiate policies across regions. Regions with more inequality or with more strongly redistributive preferences, as captured by a lower social welfare weight g^A, will prefer more progressive tax and transfer systems. However, the degree of progressivity and tax diversity is limited by the mobility of people across regions within the fiscal union (as captured by η_A^B) as well as by the mobility out of the fiscal union as a whole (as captured by η^A).

The elasticity of mobility with respect to taxes for region *A* would be smaller if (1) the region is larger (as discussed above), (2) if there is more tax coordination with jurisdictions that do not operate under the same fiscal authority, and (3) if mobility is lower due to nontax factors such as preferences and other policies. As for the latter, regulatory policies such as visa requirements and work permits, or transfer policies such as eligibility for welfare benefits and social insurance, may be important. Several countries, including France, Spain, and the United States, also impose exit or expatriation taxes for residents who decide to leave, which can be viewed as another way of trying to reduce the migration elasticity of domestic residents. Mobility responses to taxes will depend crucially on the local amenities of a region, on the public goods and services provided, and on agglomeration effects. All these forces also shape the within-jurisdiction cross-elasticity η_A^B and are plausibly even stronger within jurisdictions. Regions which are more similar in terms of amenities and thus more closely substitutable will face higher cross-elasticities and will have to set more similar tax rates than in a world without people and income mobility.

us consider the effects of a small change in the tax rate τ^A, $d\tau^A$. First, this reduces each agent's income by $y_i^A\, d\tau^A$, which costs $-g_i y_i^A\, d\tau^A$ in terms of social welfare. Aggregating across all agents, the total effect is $-\sum_i g_i y_i^A\, d\tau^A$. In addition, the mechanical effect on revenues (without agents moving regions) equals $-\sum_i g_i y^A\, d\tau^A$. Since people also move regions following the tax change, there is an additional revenue effect, equal to $-\sum_i g_i \left(\frac{\tau^A dy^A}{d(1-\tau^A)} + \frac{\tau^B dy^B}{d(1-\tau^A)}\right) d\tau^A$. Let $\eta^A \equiv \frac{dy^A}{y^A}\frac{(1-\tau^A)}{d(1-\tau^A)} > 0$ be the elasticity of income in region *A* to the net-of-tax rate $1-\tau^A$ in the region and $\eta_A^B \equiv \frac{dy^B}{y^B}\frac{(1-\tau^A)}{d(1-\tau^A)} < 0$ be the cross-elasticity of income in region *B* to the net-of-tax rate $1-\tau^A$ in region *A*. The term $g^A \equiv \frac{\sum_{i\in A} g_i y_i^A}{\sum_i g_i y^A}$ is the average, income-weighted welfare weight in region *A*. Setting the three effects to zero, rearranging, and using the definitions in the text yields the formula in the text.

If policies are instead constrained to be uniform across the two regions within the jurisdiction, then the ability to differentiate policies and adapt them to local conditions is thus lost.

As discussed above, when considering tax policy setting across independent jurisdictions (states or countries), we do not immediately see evidence of a race to the bottom. This suggests that some implicit coordination is taking place, perhaps because of a fear of retaliation along the tax policy or other margins. On the other hand, the preferential tax schemes to foreigners implemented in several countries may hint at a slippery slope towards beggar-thy-neighbor policies. Getting rid of such schemes would be a limited form of policy coordination that seems welfare-increasing in our framework and potentially feasible. Partial coordination which internalizes some, even if not all, of the welfare gains from full coordination is an intermediate solution and already exists between state and local jurisdictions in the United States and subnational jurisdictions in other countries. Examples include revenue sharing and matching or categorical grants, partially centralized provision of public goods, central tax deductibility of local government taxes, or regulations concerning what sort of taxes and tax bases may be used by local governments.

Conclusion

There is growing evidence that taxes can affect the geographic location of people both within and across countries. This migration channel creates another efficiency cost of taxation with which policymakers need to contend when setting tax policy. At the same time, we have cautioned against overusing these empirical findings to argue in favor of an ineluctable reduction in the level of taxation or progressivity. Let us reiterate two key caveats.

First, while the mobility responses documented in some of the recent literature are striking and perhaps surprisingly large, they pertain to specific groups of people and to specific countries. Although we are far from having to rely on the celebrity anecdotes presented in the introduction, data limitations and identification challenges have forced researchers to study the migration flows in specific countries (like Denmark) or to focus on a specific population internationally (like superstar football players or inventors). We are still lacking systematic evidence on the mobility elasticities of the broader population and across different types of countries.

Second, the strength of the mobility response to taxes is not an exogenous, structural entity. It depends critically on the size of the tax jurisdiction, the extent of international or subnational tax coordination, and the prevalence of other forces that foster or limit the movement of people, all of which can also be affected by policies. These forces include local or national amenities, agglomeration effects, and the provision of public goods and services. Rather than compromising redistribution or restraining free mobility in an inefficient way, these can, in a productive way, be fostered to make the country or state attractive to people.

References

Agersnap, Ole, Amalie Sofie Jensen, and Henrik Kleven. 2019. "The Welfare Magnet Hypothesis: Evidence from an Immigrant Welfare Scheme in Denmark." NBER Working Paper 26454.

Agrawal, David R., and Dirk Foremny. 2019. "Relocation of the Rich: Migration in Response to Top Tax Rate Changes from Spanish Reforms." *Review of Economics and Statistics* 101 (2): 214–32.

Akcigit, Ufuk, Salomé Baslandze, and Stefanie Stantcheva. 2016. "Taxation and the International Mobility of Inventors." *American Economic Review* 106 (10): 2930–81.

Akcigit, Ufuk, John Grigsby, Tom Nicholas, and Stefanie Stantcheva. 2018. "Taxation and Innovation in the 20th Century." NBER Working Paper 24982.

Alesina, Alberto, and Enrico Spolaore. 1997. "On the Number and Size of Nations." *Quarterly Journal of Economics* 112 (4): 1027–56.

Alesina, Alberto, Armando Miano, and Stefanie Stantcheva. 2018. "Immigration and Redistribution." NBER Working Paper 24733.

Alstadsæter, Annette, Niels Johannesen, and Gabriel Zucman. 2019. "Tax Evasion and Inequality." *American Economic Review* 109 (6): 2073–2103.

Bakija, Jon, and Joel Slemrod. 2004. "Do the Rich Flee from High State Taxes? Evidence from Federal Estate Tax Returns." NBER Working Paper 10645.

Bhagwati, Jagdish, and John D. Wilson. 1989. "Income Taxation in the Presence of International Personal Mobility: An Overview." In *Income Taxation and International Mobility*, edited by Jagdish N. Bhagwati and John Douglas Wilson, 3–39. Cambridge, MA: MIT Press.

Borjas, George J. 1999. "The Economic Analysis of Immigration." In *Handbook of Labor Economics*, Vol. 3, edited by O. Ashenfelter and D. Card, 1697–1760. Amsterdam: Elsevier Science.

Brülhart, Marius, and Raphaël Parchet. 2014. "Alleged Tax Competition: The Mysterious Death of Bequest Taxes in Switzerland." *Journal of Public Economics* 111: 63–78.

Brülhart, Marius, Jonathan Gruber, Matthias Krapf, and Kurt Schmidheiny. 2019. "Behavioral Responses to Wealth Taxes: Evidence from Switzerland." CESifo Working Paper 7908.

Conway, Karen Smith, and Jonathan C. Rork. 2006. "State 'Death' Taxes and Elderly Migration—The Chicken or the Egg?" *National Tax Journal* 59 (1): 97–128.

Feldstein, Martin, and Marian V. Wrobel. 1998. "Can State Taxes Redistribute Income?" *Journal of Public Economics* 68 (3): 369–96.

Gordon, Roger H. 1983. "An Optimal Taxation Approach to Fiscal Federalism." *Quarterly Journal of Economics* 98 (4): 567–86.

Gordon, Roger H., and James R. Hines Jr. 2002. "International Taxation." In *Handbook of Public Economics*, Vol. 4, edited by Alan J. Auerbach and Martin Feldstein, 1935–95. Amsterdam: Elsevier Science.

Griffith, Rachel, James R. Hines Jr., and Peter Birch Sørensen. 2010. "International Capital Taxation." In *Dimensions of Tax Design: The Mirrlees Review*, edited by Stuart Adam, Tim Besley, Richard Blundell, Stephen Bond, Robert Chote, Malcolm Gammie, Paul Johnson, Gareth Myles, and James M. Poterba, 914–96. Oxford: Oxford University Press.

Jakobsen, Katrine, Kristian Jakobsen, Henrik Kleven, and Gabriel Zucman. Forthcoming. "Wealth Taxation and Wealth Accumulation: Theory and Evidence from Denmark." *Quarterly Journal of Economics*.

Kanbur, Ravi, and Michael Keen. 1993. "Jeux Sans Frontières: Tax Competition and Tax Coordination When Countries Differ in Size." *American Economic Review* 83 (4): 877–92.

Keen, Michael, and Kai A. Konrad. 2013. "The Theory of International Tax Competition and Coordination." In *Handbook of Public Economics*, Vol. 5, edited by Emmanuel Saez, Martin Feldstein, Raj Chetty, and Alan J. Auerbach, 257–328. Amsterdam: Elsevier.

Kleven, Henrik J. 2018. "Sufficient Statistics Revisited." https://www.henrikkleven.com/uploads/3/7/3/1/37310663/kleven_sufficientstats_march2018.pdf.

Kleven, Henrik Jacobsen, Camille Landais, and Emmanuel Saez. 2013. "Taxation and International Migration of Superstars: Evidence from the European Football Market." *American Economic Review* 103 (5): 1892–1924.

Kleven, Henrik Jacobsen, Camille Landais, Emmanuel Saez, and Esben Schultz. 2014. "Migration and Wage Effects of Taxing Top Earners: Evidence from the Foreigners' Tax Scheme in Denmark." *Quarterly Journal of Economics* 129 (1): 333–78.

Liebig, Thomas, Patrick A. Puhani, and Alfonso Sousa-Poza. 2007. "Taxation and Internal Migration—Evidence from the Swiss Census Using Community-Level Variation in Income Tax Rates." *Journal of*

Regional Science 47 (4): 807–36.

Martinez, Isabel. 2017. "Beggar-Thy-Neighbour Tax Cuts: Mobility after a Local Income and Wealth Tax Reform in Switzerland." LISER Working Paper 2017-08.

Mirrlees, J.A. 1982. "Migration and Optimal Income Taxes." *Journal of Public Economics* 18 (3): 319–41.

Moretti, Enrico, and Daniel J. Wilson. 2017. "The Effect of State Taxes on the Geographical Location of Top Earners: Evidence from Star Scientists." *American Economic Review* 107 (7): 1858–1903.

Moretti, Enrico, and Daniel J. Wilson. 2019. "Taxing Billionaires: Estate Taxes and the Geographical Location of the Ultra-Wealthy." NBER Working Paper 26387.

Muñoz, Mathilde. 2019. "Do European Top Earners React to Labour Taxation through Migration?" WID.world Working Paper 2019/12.

Oates, Wallace E. 1972. *Fiscal Federalism.* New York: Harcourt Brace Jovanovich.

Oates, Wallace E. 1999. "An Essay on Fiscal Federalism." *Journal of Economic Literature* 37 (3): 1120–49.

Piketty, Thomas, and Emmanuel Saez. 2013. "Optimal Labor Income Taxation." In *Handbook of Public Economics*, Vol. 5, edited by Alan Auerbach, Raj Chetty, Martin Feldstein, and Emmanuel Saez, 391–474. Amsterdam: Elsevier.

Piketty, Thomas, Emmanuel Saez, and Stefanie Stantcheva. 2014. "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities." *American Economic Journal: Economic Policy* 6 (1): 230–71.

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

Rosen, Sherwin. 1979. "Wage-Based Indexes of Urban Quality of Life." In *Current Issues in Urban Economics*, edited by P. Mieszkowski and M. Straszheim, 74–104. Baltimore: Johns Hopkins University Press.

Saez, Emmanuel, and Stefanie Stantcheva. 2016. "Generalized Social Marginal Welfare Weights for Optimal Tax Theory." *American Economic Review* 106 (1): 24–45.

Saez, Emmanuel, and Stefanie Stantcheva. 2018. "A Simpler Theory of Optimal Capital Taxation." *Journal of Public Economics* 162: 120–42.

Schmidheiny, Kurt. 2006. "Income Segregation and Local Progressive Taxation: Empirical Evidence from Switzerland." *Journal of Public Economics* 90 (3): 429–58.

Schmidheiny, Kurt, and Michaela Slotwinski. 2018. "Tax-Induced Mobility: Evidence from a Foreigners' Tax Scheme in Switzerland." *Journal of Public Economics* 167: 293–324.

Tiebout, Charles M. 1956. "A Pure Theory of Local Expenditures." *Journal of Political Economy* 64 (5): 416–24.

Young, Cristobal, Charles Varner, Ithai Z. Lurie, and Richard Prisinzano. 2016. "Millionaire Migration and Taxation of the Elite: Evidence from Administrative Data." *American Sociological Review* 81 (3): 421–46.

Zucman, Gabriel. 2013. "The Missing Wealth of Nations: Are Europe and the U.S Net Debtors or Net Creditors?" *Quarterly Journal of Economics* 128 (3): 1321–64.

Journal of Economic Perspectives—Volume 34, Number 2—Spring 2020—Pages 143–171

The Separation and Reunification of Germany: Rethinking a Natural Experiment Interpretation of the Enduring Effects of Communism

Sascha O. Becker, Lukas Mergele, and Ludger Woessmann

German separation in 1949 into the German Democratic Republic and the Federal Republic of Germany and its reunification in 1990 seem to offer a unique setting of a rather unexpected introduction and termination of a communist regime in one part of a previously and afterward unified country. Analyzing East-West differences in Germany provides the opportunity to study effects of living in different political systems, which has general relevance for our understanding of the fundamentals of economic preferences and behavior. This paper shows that because of preexisting differences and early selective migration, German division and reunification do not provide a straightforward case of a natural experiment. Taking these challenges into account, it summarizes what can

■ *Sascha O. Becker is Xiaokai Yang Chair of Business and Economics, Monash University, Melbourne, Australia; Professor of Economics, University of Warwick, Coventry, United Kingdom; and Research Fellow, Centre for Economic Policy Research (CEPR), London, United Kingdom. Lukas Mergele is Postdoctoral Research Economist at the ifo Institute and a member of the Junior Faculty Program, University of Munich, both in Munich, Germany. Ludger Woessmann is Professor of Economics, University of Munich, and Director of the ifo Center for the Economics of Education, both in Munich, Germany. Becker and Woessmann are also Research Fellows, Institute for Labor Economics (IZA), Bonn, Germany; Research Network Fellows, CESifo, Munich, Germany; Research Associates, Centre for Competitive Advantage in the Global Economy (CAGE), Warwick University, United Kingdom; and Research Fellows, Research Centre for Education and the Labour Market (ROA), Maastricht University, Netherlands. Their email addresses are sascha.becker@monash.edu, mergele@ifo.de, and woessmann@ifo.de.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.2.143.

be learned from the literature that studies how individuals who lived in East and West Germany differed at reunification and how these differences have evolved.

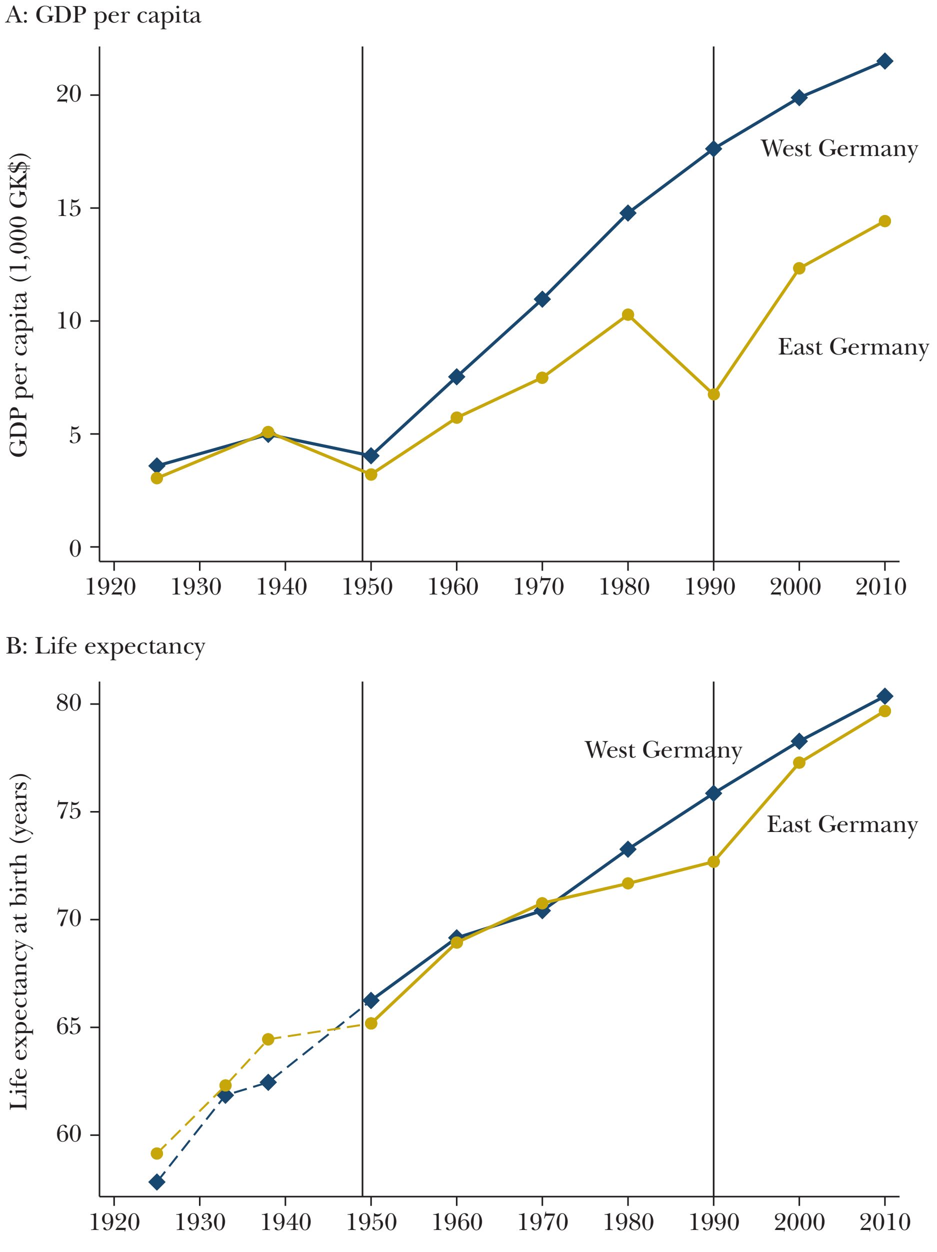
Under the communist regime, economic and general well-being of the East German population fell behind the development in West Germany, as shown in Figure 1. Of course, such estimates should be interpreted with care. GDP estimates for the communist East are not based on market prices and were potentially subject to misreporting. Simple comparisons should not be taken as a precise causal effect of communism on economic prosperity or other aspects of well-being. However, it is clear that by the end of the 1980s, people in the West were immensely better off. Maybe more surprisingly, while life expectancy in the East has converged quickly to the Western level after reunification, the convergence of economic activity has stalled after an initial surge—despite continuing major transfers from West to East. Many scholars of transition economics had expected that changing the system would suffice to ensure convergence after undoing artificial distortions that central planning imposed on the Eastern economy. Initial decisions about converting East German to West German marks at 1:1 parity, how to carry out privatization, and other policy choices may have contributed to the path of East German economic activity after reunification. However, a growing literature emphasizes that living under communism may have changed people's attitudes and preferences more deeply, giving rise to much more enduring effects even after the end of the political regime—and suggesting that transitions away from communism are about more than removing policy distortions.

Learning about the effects of communism based on the experience of German division and reunification poses unique challenges. To be considered a natural experiment, the exposure to different political systems would have to be unrelated to any other characteristics of the population that may be related to the outcomes of interest.

In the first sections of this paper, we highlight several sources of endogeneity that would violate the interpretation of the differential exposure to political systems as natural experiments. First, the drawing of the border between East and West Germany was not random. We demonstrate that substantial differences in economic structures, political preferences, cultural traits, and gender roles between what later became East and West Germany existed well before World War II. Second, East and West Germany were differentially affected by the war and by the dismantling of infrastructure and reparations to the occupying forces in the immediate after-war period. Third, roughly one-fifth of the East German population moved to West Germany between 1945 and the building of the Berlin Wall in 1961, and this out-migration was likely selective with respect to political and economic preferences. Prior literature considers some of these aspects but tends to miss important preexisting differences by using rough measures or by aggregating data that combine the area of what would later become the German Democratic Republic with areas that would be part of Poland. For this paper, we collect fine-grained data on a broad series of indicators at the county level that allow us to provide a clearer picture of preexisting differences. Together, this evidence suggests that we might expect

Figure 1

GDP per Capita and Life Expectancy in East and West Germany, 1925–2010



A: GDP per capita

B: Life expectancy

Source: GDP per capita: Rosés and Wolf (2018), own calculations. Pre-1950 life expectancy: Wagner (2008), own calculations. Life expectancy from 1950 onwards: Luy (2020).

Note: GDP data are in 1990 Geary-Khamis US dollars and exclude Berlin. Pre-1950 life expectancy data comprise the states of Baden, Bavaria, Hesse, and Wurttemberg (equi-weighted average) for West Germany and Thuringia and Saxony for East Germany. From 1950 onwards, life expectancy data comprise the respective entire territories.Vertical lines represent German separation (1949) and reunification (1990).

substantial differences in attitudes between individuals who lived in East and in West Germany even in the absence of differential exposure to political systems.

We then turn to an overview of the available evidence about Germany's experience of separation and reunification on enduring differences in economic outcomes, political attitudes, cultural traits, and gender roles. We do not attempt a formal reworking of any of this literature but instead focus on whether the central results of various studies are likely to have an upward or downward bias when aspects of preexisting historical differences, differential treatment during World War II and the occupation period, and German-German migration before the Berlin Wall are taken into account.

Our discussion also suggests a broader lesson beyond the effects of German reunification. There is a wide literature exploring how political systems persistently affect the economy and population preferences, with a particular focus on different legacies between capitalist and socialist societies (for a review, see Simpser, Slater, and Wittenberg 2018). Studies examining a major political event, like the arrival or departure of communism or other forms of autocracy, might face similar challenges when interpreting the before-and-after differences as a natural experiment. The ways in which borders were originally drawn, events that took place during the transition, and patterns of in- or out-migration are often important. Thus, we are proposing the need for a reinterpretation of the existing literature on the "effects" of communism and we call for a closer consideration of the *formation* of political systems more generally.

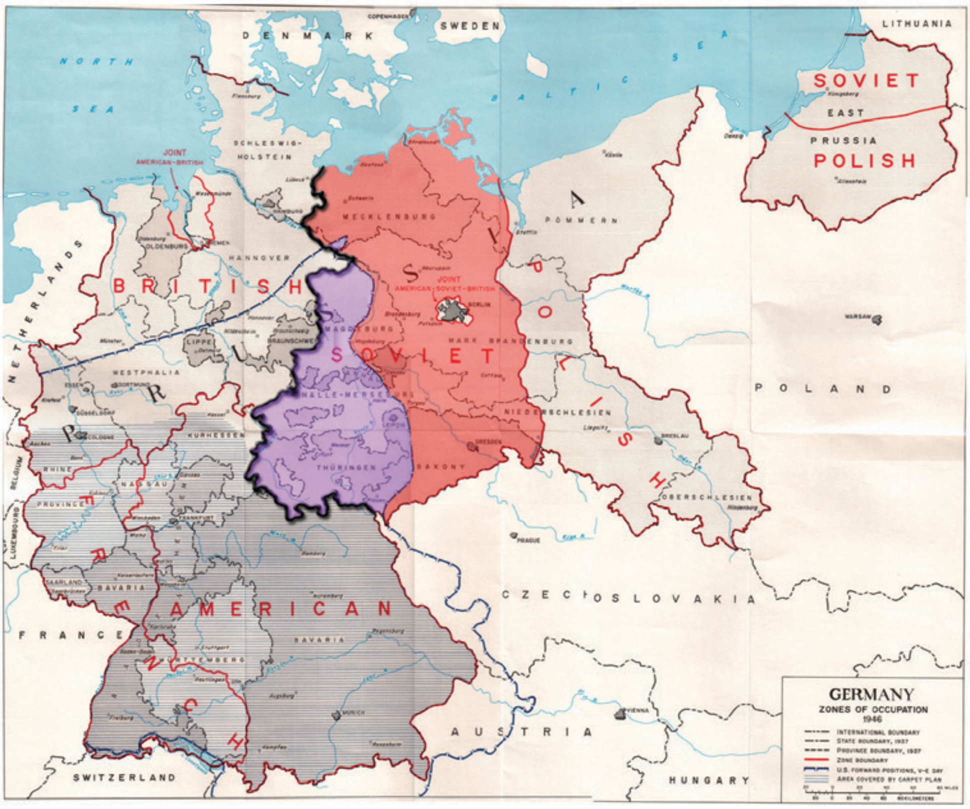
Preexisting Differences between East and West Germany

The location of the border between the German Democratic Republic (GDR) and the Federal Republic of Germany (FRG) is *not* the random outcome of where American, British, and Soviet tanks stopped at the end of World War II in 1945. Instead, in anticipation of the defeat of Nazi Germany, the foreign ministers of the United States, United Kingdom, and Soviet Union agreed on the formation of the European Advisory Commission at the Moscow Conference on October 30, 1943, which was confirmed at the Tehran Conference in November. The commission was asked to explore the political framework for postwar Europe and to make recommendations to the three governments (Kowalski 1971).

Based on the recommendations of the European Advisory Commission, the post-World War II border between East and West Germany was agreed upon in the so-called London Protocol, signed on September 12, 1944. The American, British, and Soviet armies were each allocated a zone of occupation of roughly equal population size. Berlin was to be jointly occupied. Two changes were made later: First, an additional French zone was carved out of the American and British zones. Second, parts of the Soviet zone were given to Poland (so that the Soviet-Polish border followed the Oder-Neisse line), and Poland became a Soviet satellite country. Neither change affected the German-German border, which separated the Soviet zone from those of the Western allies.

Figure 2

German Post-World War II Occupation Zones, Pre-World War II Provinces, and End-of-War Frontline



Source: Earl F. Ziemke, *The U.S. Army in the Occupation of Germany*, 1975. Library of Congress, Catalog Card Number 75-619027.
Note: Occupation zones in post-war Germany, highlighting the Soviet zone (red and purple), the inner German border (heavy black line), and the zone from which American forces withdrew in July 1945 (purple). With minimal exceptions, the ultimate border between the German Democratic Republic and the Federal Republic of Germany in the years 1949–1990 follows the pre-World War II province borders, the border proposed in the EAC protocols in 1944 and 1945 as well as the Western border of the Soviet occupation zone from 1945–1949 (depicted).

Figure 2 shows that the border between the Soviet and the Western occupation zones followed the pre-World War II borders of states of the German Empire and provinces of the largest state of Prussia (with a few very minor exceptions for geographic connectedness). As a result, the German-German border separated the populations of preexisting regions with distinct histories and cultures. The British delegation to the European Advisory Commission argued that this approach tended to “encourage rather than [. . .] prejudice any separatist or particularist tendencies,” providing a safeguard against a revival of the former German state (US Department of State 1968, pp. 150ff). The border thus was entirely different from the end-of-war frontline between the Red Army and the US and British armies. Indeed, by May 1945, the Western armies had already pushed deeply into what later became East German territory but handed over control to the Soviets once German defeat was accomplished, thereby honoring the agreement made in the London Protocol. In 1949, the Soviet occupation zone became the German Democratic Republic, and the Allied occupation zones became the Federal Republic of Germany.

Given that the border between East and West Germany was a deliberate choice and followed preexisting provincial boundaries, the possibility arises that populations living in the two parts may have differed before the advent of communism in the East. Analyzing pre-World War II data for counties of what would later become the German Democratic Republic and the Federal Republic of Germany, we find important preexisting East-West differences in several measures of economic outcomes, political preferences, culture, and gender roles.

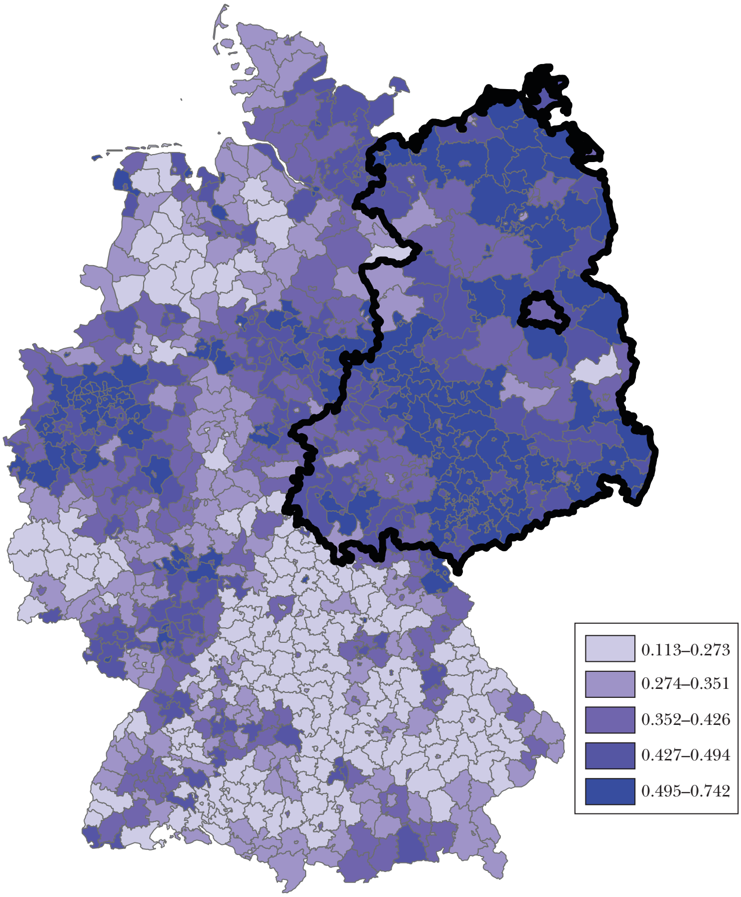
Economic Outcomes

To investigate whether East-West differences in economic structures predate the division into East and West Germany, we draw on data from the 1925 German Census compiled by Falter and Hänisch (1990; see also Hänisch 1989). Restricting the analysis to what would later be East and West Germany, the county-level data allow us to compare 202 counties in the East to 652 counties in the West. Figure 3 illustrates the working-class structure of the Eastern population: The East-West border is apparent already before World War II, with the working-class share substantially higher in what would later become the communist East.

Using these data in regression analyses, the first entry in the first row of Table 1 shows that the share of blue-collar workers in the total labor force in the West was 35.6 percent.[1] The second row shows that this share was 11.8 percentage points higher in the counties that would later become part of East Germany. In addition, the working-class share jumps abruptly in several regions around the later inner-German border, especially in the southwest and the south of what later becomes East Germany. As shown in the remaining columns of Table 1, the difference in working-class shares is just as apparent when focusing on the 468 counties within 200 kilometers of the later border or the 244 counties within 100 kilometers of the later border (analyses that also entail a more balanced number of counties on either side of the border). A difference of 7.3 percentage points even shows up

[1] The numbers in Table 1 are based on a series of separate regressions, using county-level data. For details on these regressions, see the online Appendix available with this paper at the *JEP* website. For additional results and more detail on preexisting differences between East and West Germany, see Table A1 and Figures A1–A4 in the online Appendix. While the inner-German border provides a clean separation of pre-World War II counties into East and West Germany (as discussed above), several pre-World War II counties straddled across both sides of the Oder and Neisse Rivers that formed the later GDR-Polish border. We include all those counties in our analysis whose county capital is part of the later East Germany. Berlin is excluded from the analysis throughout.

Figure 3
The Working-class Share in 1925: East-West Differences before the GDR Existed



Source: Own depiction based on county-level census data in Falter and Hänisch (1990). Base map: MPIDR (2011).
Note: Colors refer to quintiles. Missing data imputed by nearest neighbor.

when narrowing the analysis to only those 59 counties that have a direct contact with the later border.

The difference in occupational structure is also reflected in the sectoral composition of the economy, where the employment share in manufacturing (industry and crafts) is 8.3 percentage points larger in the Eastern compared to Western counties of Germany, where it was 35.7 percentage points. When restricting the analysis to counties within 100 or 200 kilometers of the later border, the difference is even larger at more than 11 percentage points. The larger manufacturing share in what would become East Germany mirrors the reverse pattern in the agricultural sector. While the agricultural employment share was 35.2 percent in the West, it was only

Table 1
East-West Differences before World War II

	Full German sample	Band around the East-West border		
		<= 200 km	<= 100 km	Border counties
Economic outcomes				
Working-class share (1925)				
Constant (West)	0.356***	0.340***	0.363***	0.369***
	(0.005)	(0.006)	(0.009)	(0.020)
GDR	0.118***	0.136***	0.108***	0.073***
	(0.007)	(0.009)	(0.012)	(0.024)
Manufacturing employment share (1925)				
Constant (West)	0.357***	0.332***	0.349***	0.341***
	(0.006)	(0.008)	(0.012)	(0.028)
GDR	0.083***	0.121***	0.110***	0.031
	(0.012)	(0.015)	(0.019)	(0.039)
Self-employment share (1925)				
Constant (West)	0.273***	0.281***	0.263***	0.269***
	(0.004)	(0.005)	(0.007)	(0.014)
GDR	–0.084***	–0.093***	–0.068***	–0.039**
	(0.005)	(0.006)	(0.008)	(0.017)
Political preferences				
Communist (KPD) vote share (1924)				
Constant (West)	0.050***	0.037***	0.036***	0.035***
	(0.002)	(0.002)	(0.003)	(0.004)
GDR	0.049***	0.069***	0.068***	0.021*
	(0.005)	(0.006)	(0.008)	(0.011)
Left (SPD+USPD+KPD) vote share (1924)				
Constant (West)	0.254***	0.274***	0.307***	0.316***
	(0.005)	(0.008)	(0.011)	(0.027)
GDR	0.152***	0.138***	0.100***	0.028
	(0.008)	(0.011)	(0.016)	(0.037)
Culture				
Protestant share (1925)				
Constant (West)	0.491***	0.602***	0.663***	0.754***
	(0.015)	(0.021)	(0.028)	(0.051)
GDR	0.420***	0.307***	0.240***	0.099
	(0.016)	(0.023)	(0.031)	(0.068)
Church attendance (1900-1910)				
Constant (West)	0.555***	0.653***	0.627***	0.638***
	(0.016)	(0.021)	(0.032)	(0.066)
GDR	–0.165***	–0.269***	–0.224***	–0.199**
	(0.019)	(0.024)	(0.036)	(0.075)
Gender roles				
Female labor-force participation (1925)				
Constant (West)	0.325***	0.324***	0.326***	0.308***
	(0.003)	(0.004)	(0.007)	(0.011)
GDR	0.060***	0.058***	0.049***	0.037**
	(0.006)	(0.007)	(0.010)	(0.016)
Extramarital birth ratio (1937)				
Constant (West)	0.067***	0.070***	0.069***	0.065***
	(0.002)	(0.002)	(0.002)	(0.005)
GDR	0.033***	0.030***	0.025***	0.014*
	(0.002)	(0.003)	(0.004)	(0.008)
Observations (vote shares)	854	468	244	59

Note: Each pair of "Constant (West)" and "GDR" estimates reflects the result of a separate regression. County-level analyses. Distance to border based on county centroids. Robust standard errors in parentheses: significance at * 10, ** 5, *** 1 percent. Data sources: economic outcomes, political preferences, and Protestant share: Falter and Hänisch (1990); Protestant church attendance: Hölscher (2001); female labor-force participation: Wyrwich (2019); extramarital birth ratio: Klüsener and Goldstein (2016). For additional details, see Appendix Table A1 in the online Appendix.

22.4 percent in the East. By contrast, there are hardly any East-West differences in average employment shares in sectors reflecting basic services such as administration, domestic services, and the health sector. For example, health-sector employment shares do not differ significantly between counties in the East and the West.

Given the important difference between planned and market economies in allowing for entrepreneurship and self-employment, the literature on contemporaneous East-West differences places a strong focus on these outcomes. However, the self-employment share was already substantially lower in the East than in the West in 1925 (Fritsch and Wyrwich 2014). As shown in Table 1, the difference is 8.4 percentage points on average and endures when focusing on counties closer to the later border.

These pervasive differences in economic structure are consistent with an analysis of trade flows between different parts of the country. Wolf (2009) shows that the German Empire was a poorly integrated economy before 1914 and was only "reasonably well integrated" by the end of the Weimar Republic.

Political Preferences

To analyze preexisting political preferences, a natural starting point is to look at vote shares for the Communist Party of Germany (*Kommunistische Partei Deutschlands*, KPD) before World War II. During the Weimar Republic (1918–1933), regular elections for the Reichstag were held throughout Germany. We study county-level data on vote shares of different parties in the national Reichstag election of December 1924.[2] The second panel of Table 1 illustrates that counties that would later become part of East Germany have a communist vote share that is 4.9 percentage points higher compared to the West's 5.0 percent, and thus nearly twice as high. This difference is even larger, at close to 7 percentage points, when focusing on the counties within 100 or 200 kilometers of the later border.

The East-West difference is even larger when considering all left-wing parties, which include the much larger Social Democratic Party (*Sozialdemokratische Partei Deutschlands*, SPD) in addition to the KPD as well as the negligible USPD. In the West, 25.4 percent of voters supported these parties, while the left-party vote share was 15.2 percentage points larger in the East, totaling 40.6 percent. The stronger support for communist and left-wing parties is likely linked to the larger working-class and manufacturing shares indicated above.

Clearly, studies looking at Germany's separation and reunification cannot implicitly assume that the regions—or even the areas relatively close to the border—were originally similar in their political leanings. Results for vote shares of other political parties indicate that the larger preferences for left-wing parties in the East come largely at the detriment of vote shares for the Catholic center-right *Zentrum*

[2] The December 1924 election provides an informative example as it took place after the establishment of the Weimar electoral system but before the rise of National Socialism, which disrupted the existing party system by becoming the first "mass-integration party" (for example, Falter 2014). Still, we find qualitatively similar results for other elections such as May 1924 and May 1928.

party. This relates to the fact that the *Zentrum* party was the focal party of Catholics in Germany, and since counties in the East were predominantly Protestant (as discussed next), *Zentrum* had lower vote shares in those areas. The electoral data also reveal that East and West varied in voter turnout, which was 5 percentage points higher in the East. This difference disappears, however, when restricting the analysis to counties within 100 kilometers of the later East German border.

Culture

One dimension of cultural differences that is available in historical census data is religious denomination. The share of Protestants was higher in the East than in the West, which partly reflects that Martin Luther's city of Wittenberg is situated in the middle of what would become East Germany (Becker and Woessmann 2009). In the 1925 German census, 91 percent of the population in the Eastern counties was Protestant compared to only 49 percent in the Western counties, as shown in the third panel of Table 1. The difference becomes smaller when zooming in on counties situated closer to one another, but it is still 24 percentage points in the sample of counties within 100 kilometers of what later became the German-German border.[3]

Echoing the larger Protestant share, the share of Catholics in the East was 43.8 percentage points smaller, compared to the 49.1 percent in the West. The share of Jewish population was slightly but significantly lower in the East, albeit at a very low level of 0.2 percent compared to 0.5 percent in the West. However, Berlin—which had by far the largest Jewish community in the German Empire but was divided between East and West—is excluded in this analysis.

Beyond denominational affiliation, statistical surveys of the Protestant Regional Churches of Germany on the "Expressions of Churchly Life" provide a historically unique indicator of church attendance. Based on headcounts combined in Sacrament Statistics (*Abendmahlsstatistik*), the measure refers to the number of participations in Holy Communion relative to the number of Protestants, which was used by contemporaries as a proxy for churchliness (see also Becker and Woessmann 2013, 2018; Becker, Nagler, and Woessmann 2017). We follow Hölscher (2001), who gathered the data at the level of church districts (*Kirchenkreise*) from regional archives and focuses on the year 1910 because of broad data coverage, but we use the average of the data available for the years 1900–1910 to reduce measurement error.

Church attendance at this time was substantially lower to the East of what later became the East German border. As shown in Table 1, participations in Holy Communion were 16.3 percentage points lower among East German Protestants, compared to an average of 55.4 percent among West German Protestants. This difference is robust in the smaller bands around the border and holds even for the counties contiguous to the border. Hölscher (2001, p. 7) notes: "A look at the map of Protestant Communion participation in 1910. . .already reveals the later German

[3] Grashoff (2019) notes that suicide rates during the Weimar Republic were also higher in the East, which he sees as a corollary of more widespread Protestantism (Becker and Woessmann 2018).

dividing border before the First World War between Hesse and Thuringia. This suggests the conjecture that [. . .] in large parts of East Germany it was not the socialist regime that first eroded and undermined ecclesiastical life, but rather that, conversely, an already older unchurchliness in these regions paved the way for the reception of socialist [. . .] convictions."

Gender Roles

With respect to gender roles in the labor market, the 1925 census provides employment statistics by gender. Wyrwich (2019) presents county-level data on female labor-force participation, measured as the share of women registered as nondomestic employees within the entire female population. As pointed out in Wyrwich (2017), participation of females in the formal labor market was higher on average in East relative to West Germany before World War II. At the 1925 county level, female labor-force participation was 6.0 percentage points higher in the East compared to the mean of Western counties of 32.5 percent, as shown in the fourth panel of Table 1. Again, the significant difference also emerges in the more geographically restricted samples.

In addition, Klüsener and Goldstein (2016) have shown that East-West differences in family-formation behavior, as documented in extramarital births, predate the 1945 division of Germany. They provide data on extramarital birth ratios in 1937 published by the German Imperial Statistical Office, confirming that extramarital fertility was higher on average in the East. As shown in Table 1, the difference amounts to 3.3 percentage points on average, compared to the Western mean of 6.7 percent. This difference is also evident in the restricted samples.

Overall, our analysis documents remarkable historical differences in economic outcomes, political preferences, culture, and gender roles between the populations living in the regions that were to become East and West Germany. Well before World War II, people in the later East Germany were more likely to be working class and to work in manufacturing, less likely to be self-employed, and more favorable to communist and left-wing political parties. The East had higher Protestant shares, lower church attendance, higher female labor-force participation, and higher extramarital fertility. To the extent that such preexisting differences persist through the communist period, they may well be an essential source of post-reunification heterogeneities between East and West Germans.

Differential Affectedness by World War II and Occupying Forces

A further source of East-West differences is the potentially differential effect of World War II itself on the different parts of Germany as well as potential differences in the treatment received during the years 1945–1949, when East and West Germany were occupied by Soviet and Allied Western armies, respectively. Economic historians have long noted differences in labor productivity in manufacturing between East and West Germany predating World War II. Using the 1936 Manufacturing

Census, van Ark (1996) shows that sales per employee in East Germany amounted to only 84 percent of the level in West Germany.[4]

In his comprehensive comparison of the East German and West German economies, Sleifer (2006) notes that after 1944, the East and West German industrial capital stocks showed a strongly divergent development. This is *not* the result of larger damage due to war activity but largely attributed to the dismantling of East Germany's industrial capital stock by the Soviet Union. Baar, Karlsch, and Matschke (1995) estimate that East Germans paid much more than West Germans (2,784 versus 1,611 Reichsmarks per capita) in terms of total war damages, dismantling, reparations, and occupation costs (also see Sleifer 2006, Table 4.3). The difference is mainly due to higher reparations (1,065 versus 23 Reichsmarks per capita) and higher losses due to dismantling of capital equipment (384 versus 60 Reichsmarks per capita) in the East. War damages are actually lower in East Germany compared to the West (686 versus 839 Reichsmarks per capita), and the costs of occupation are roughly similar (649 versus 689 Reichsmarks per capita).

Overall, while the West German industrial capital stock in 1948 was higher than in 1936, the East German capital stock was at only 69 percent of its 1936 level (Sleifer 2006). Considering that the East German manufacturing sector was already at a disadvantage in 1936, this means that it had fallen significantly further behind by the time the German Democratic Republic was established.

Another indication of differences arising between East and West already before the German Democratic Republic and the Federal Republic of Germany were officially founded in 1949 can be seen when looking at sex ratios of men to women. We digitized county-level data from the German Census jointly administered in all four occupation zones on October 29, 1946. The sex ratios in the four occupation zones (excluding Berlin) were 0.820 in the American zone, 0.835 in the British zone, 0.790 in the French zone, and 0.743 in the Soviet zone. There had been no such differences in 1939 in the last pre-World War II census when sex ratios varied only between 0.954 and 0.974 across the four areas. The larger decrease in sex ratios in the Soviet zone may reflect a combination of a larger fraction of war casualties as well as sex imbalances in very early East-West migration. Whatever the source, considering the well-known implications of imbalances in sex ratios for labor-market outcomes (for example, Angrist 2002), these differences might have contributed to differences in several outcomes of interest, such as female labor-force participation, gender roles, and even political outcomes.

Overall, war-related damages and differences between occupying forces in the Soviet and non-Soviet zones implied that East Germany was off to a worse start even before the new states had a chance to develop their own identities.

[4] This lower productivity of the manufacturing sector counteracted the larger share of manufacturing in the East. Pre-World War II GDP per capita was not very different between West and East Germany (see Figure 1 above).

Selective Out-migration before the Berlin Wall

The Soviet Occupation Zone was established in East Germany right after the end of World War II in 1945, culminating in the foundation of the German Democratic Republic in 1949. Although the freedom of movement was restricted, the "Iron Curtain" was at first by no means impenetrable. Throughout the 1950s, people could move rather freely between the East and West sectors of Berlin, resulting in substantial East-West migration. These movements only ended with the construction of the Berlin Wall in 1961.

Statistics about migration from the Soviet Occupation Zone to the Western Occupation Zones during the years 1945–1949 are considered somewhat problematic. Plausible estimates from Heidemeyer (1994, Table 2) suggest that about 875,000 residents of the Soviet Occupation Zone moved to the Western Occupation Zones during the years 1944/45–1949. Statistics for the years 1950–1961 are considered more reliable. Van Melis (2006, Table 1) presents monthly statistics on migration from East Germany into West Berlin and West Germany between September 1950 and December 1961 that add up to 2.75 million East-West migrants. Comparing the estimated East-West migration over the entire 1945–1961 period to the roughly 18 million inhabitants in the Soviet Occupation Zone in 1946, about one-fifth of East Germany's population migrated West until 1961 when the Iron Curtain was completed and East-West migration was all but shut down.

Evidence on the Selectivity of East-to-West Migration

The evidence on the selectivity of this East-West migration is suggestive, but not conclusive. Economic research suggests in general that migrants tend to be people who are more willing to take matters into their own hands than stayers, more entrepreneurial, and selected along other dimensions (for example, Borjas 1987; Grogger and Hanson 2011; Fairlie and Lofstrom 2015; Parey et al. 2017). More specifically, it seems likely that those who left the East for the West had less preference for a communist system and were more supportive of a capitalist one, on average. This applies to ordinary citizens and leading politicians alike. In 1945, the Christian-Democratic Union (CDU)—the party of Konrad Adenauer and Angela Merkel—was founded across all zones of occupation, including the Soviet one. In the first years after 1945, the CDU had several ministers in zone-wide or state-level governments in the East. However, all CDU ministers resigned (voluntarily or by force) and migrated to the West. The Soviet occupation forces and the later East German government also expropriated many large landowners and used de-Nazification to expel not only Nazis but also those they perceived as critical to communist rule (Jessen 1999).

Using West German datasets, we can compare some characteristics of those who moved from the East to the West with people who had always been in the West (see also Bauernschuster et al. 2012). Data from a retrospective survey in the German Microcensus 1971, which covers a representative 1 percent sample of the German population, allow us to compare characteristics of those individuals

Table 2

Comparison of East-West Movers and Local West Germans in 1939: Retrospective Evidence

	East-West movers	*Local West Germans*
Occupation in 1939		
Unskilled worker	0.154	0.301
Farmer (self-employed)	0.057	0.053
Family worker	0.094	0.113
Skilled worker	0.155	0.196
White-collar worker	0.304	0.203
Civil servant	0.135	0.061
Self-employed	0.101	0.072
Education		
Basic school (*Volksschule*)	0.392	0.608
Secondary and professional school	0.490	0.348
High school	0.042	0.014
Technical school	0.027	0.012
University	0.049	0.019
Observations	**2,288**	**104,128**

Note: Retrospective evidence from German Microcensus 1971 (Handl, Mayer, and Müller 1975). Shares in sample population. Individuals still in education or outside the labor force in 1939 as well as expellees from formerly German territories excluded.

who left the Soviet Occupation Zone and then the GDR in its early days with local West Germans. Table 2 shows that early East-West movers were more likely than local West Germans to be white-collar workers (30.4 versus 20.3 percent), civil servants (13.5 versus 6.1 percent), or self-employed (10.1 versus 7.2 percent).[5] There is related selection on education: early East-West movers were much more likely to have more than basic schooling (60.8 percent) than local West Germans (39.2 percent). Furthermore, Bauernschuster et al. (2012) present evidence from the German General Social Survey (ALLBUS) 1991–2004 showing that those who had left the East for the West before 1961 see a lesser role for the state than the original West Germans.

There is also some anecdotal evidence for selection on age, health, and family status. The president of the Statistical Office of the Soviet Occupation Zone (SBZ), Bruno Gleitze, remarked that "the Soviet Occupation Zone acted like a sieve, holding back the aged, sick and single" (*"die SBZ [wirkte] wie ein Sieb, das Alte, Kranke und Alleinstehende zurückhielt,"* cited according to Steiner 2013, p. 14).

Recently, Eder and Halla (2018) argue that a substantial part of East-West migration towards the end of World War II was due to concerns about the advancing Soviet Army. They suggest that the dominant motive of migration in those

[5] We are grateful to Oliver Falck for extensive support in producing this table.

years was escaping physical assault by the Soviet army, not avoiding the socialist regime. Their evidence shows that these migrants are strongly positively selected on skills.

Together, these analyses indicate that individuals staying in East Germany differed from those who moved to the West in being less likely to be white-collar workers or self-employed, less educated, and probably more receptive to the communist doctrine. Furthermore, the large majority of emigrants from East Germany went to live in West Germany, as opposed to moving to a different country outside Germany, thus becoming part of the West in East-West comparisons. Therefore, any post-reunification differences observed between people who lived in the East and people who lived in the West may also be the result of this earlier pattern of selection, rather than just of living under a communist system in the German Democratic Republic.

A Note on Selective West-to-East Migration

About half a million people migrated from the West to the East before 1961. One prominent example is German chancellor Angela Merkel, who was born in Hamburg in 1954 and moved to East Germany as an infant when her father, a Lutheran clergyman, received a pastorate in Brandenburg. While there is no reliable data on the composition of the West-East migration, it seems likely that—in addition to return migrants—mostly individuals committed to (or at least tolerant of) the communist idea would have moved this way. Propaganda from the East German government named various motives for such migration, including young men trying to escape compulsory military service in the West or those not in agreement with the capitalist system.

We collected data showing that a considerable share of the Politburo members in the early German Democratic Republic had been born in the West (Müller-Enbergs et al. 2010). In the years 1949–1961—the years between the foundation of the German Democratic Republic and the construction of the Berlin Wall—the East German Politburo had 19 members in total. Of these, 10 were born in the territory of what later became East German area or in Berlin, three were born in areas outside the later East or West Germany, and six were born in what became West Germany. While some of these Politburo members had lived in Berlin before World War II, they deliberately selected into building the East German state. As a prominent example, Erich Honecker, born in the Saar area near the French border in what later became West Germany, joined the Politburo in 1958 and was the last leader of the German Democratic Republic when the Berlin Wall came down in 1989. Some of the most strongly convinced communists in the East came from the West.

Re-interpreting Evidence on the Effects of Communism in Germany

In light of our evidence for preexisting East-West differences, differential effects of World War II and subsequent occupying forces, and selective East-West migration, the question arises whether the previous literature potentially over- or underestimates the effects of communism when this history is taken into account. Generally speaking, in cases where some of the differences in post-1989 outcomes reflect preexisting differences, the true effect of living in East Germany will be smaller than previous studies suggest. Conversely, in cases where preexisting differences are of opposite sign to the ones found after communism, the true effect of communism may be even larger. When considering migration, it is important to ask whether it is selective with respect to the outcome of interest because not all outcomes will be equally affected by concerns of selective migration.[6] While encompassing analyses of these issues are generally missing in the literature, some papers make an explicit effort to take them into account.

In this section, we look at some of the rapidly expanding literature on German division and reunification, focusing again on the four domains of economic outcomes, political preferences, culture, and gender. Table 3 summarizes the papers we cover in terms of data, empirical approach, and results. We are not formally re-analyzing the previous work but instead discuss how the interpretation of results might have to be adjusted in light of our findings of preexisting differences and selective migration.

Economic Outcomes

In terms of overall economic outcomes, GDP and income per capita did not differ widely between East and West Germany before World War II, as shown earlier in Figure 1 (see also Alesina and Fuchs-Schündeln 2007). By the time the German Democratic Republic collapsed, its GDP per capita was less than half of that of West Germany (Sleifer 2006, Graph 3.1). After reunification, labor productivity in East Germany was at one-third of the Western level, putting the East somewhere between Mexico and Chile. Most of the capital stock of the former East Germany was obsolete or unusable for production in a market economy (Siebert 1991; Akerlof et al. 1991). The communist experiment had ended in economic failure.

How did the communist experience affect subsequent economic behavior of the people who had lived under the communist system? The recent literature exploring German division and reunification covers a wide range of economic measures, including entrepreneurship, job satisfaction, stock-market participation, savings behavior, and inflation expectations. It is hard to draw sweeping conclusions across this range of outcomes, but we will point to some enduring effects of communism that are more robust than others to the caveats we describe in the earlier sections. We start with areas of direct effects of communism and then move to topics focused on developments over the phase after reunification.

One question is whether the centrally planned economy took away the spirit of entrepreneurship. Self-employment (which typically overlaps with entrepreneurialism) was highly restricted under communism. However, post-reunification

[6] Arguably, migration unleashed by Soviet occupation or the foundation of East Germany could be defined as part of a more broadly defined communist "treatment."

Table 3

Evidence on East-West Differences

Paper, Outcome	*Data*	*Empirical approach*	*Result*
Economic outcomes			
Fuchs-Schündeln and Schündeln (2005): Saving	SOEP 1992–2000	East-West comparison: GDR versus FRG resident; civil servants versus other occupations	Precautionary wealth as share of total wealth in East amounts to twice that in West after reunification
Fuchs-Schündeln (2008): Saving	SOEP 1992–2000	East-West comparison: GDR versus FRG resident before 1990	East Germans have higher saving rates than West Germans; East-West gap increasing in age at reunification; per cohort, gap declining over time
Fuchs-Schündeln (2009): Job satisfaction	SOEP 1990–2000	Sample of East Germans: self-employed versus non-self-employed	Self-employment, mostly prohibited in GDR, possible after reunification; self-employed in East report higher job satisfaction than the employed, even controlling for income and hours worked
Fritsch and Wyrwich (2014): Entrepreneurship	Regional data 1925, 1984–2005	East-West comparison over nearly 100 years	Entrepreneurship rates highly persistent from pre-World War II to today; West Germany had higher entrepreneurship already before World War II
Laudenbach, Malmendier, and Niessen-Ruenzi (2020): Stock-market participation	Retail investor accounts of a broker, 2004–2012	East-West comparison and treatment intensity (“communist cities”)	East Germans invest less in stock market, more likely to hold stocks of companies in communist countries (China, Russia, Vietnam); effects stronger when exposed to communist priming
Goldfayn-Frank and Wohlfart (forthcoming): Inflation expectation	Bundesbank Panel of Household Finances (PHF) 2011, 2014; GfK Consumer Climate Survey 2000–2016; Eurobarometer 2000–2016	East-West comparison	East Germans expect higher inflation than West Germans
Political preferences			
Alesina and Fuchs-Schündeln (2007): Preferences for state intervention	SOEP 1996-2002	East-West comparison: GDR versus FRG residents before 1990	East Germans are more in favor of state intervention in social services, insurance, and redistribution; stronger for older cohorts; slow convergence
Svallfors (2010): Attitudes to state intervention	International Social Survey Program (ISSP), Role of Government, 1990, 1996, 2006	East-West comparison, focus on convergence	Considerable convergence in attitudes between East and West Germany: Attitudes in West Germany are completely stable while attitudes in the East become more similar to those in the West
Brosig-Koch et al. (2011): Solidarity	Laboratory experiments in Magdeburg and Essen, 1995 and 2009	East-West comparison, focus on convergence	East Germans show consistently less fairness and willingness to cooperate in solidarity games; there has been no convergence in the 20 years after the reunification
Avdeenko (2018): Voting for socialist party	aggregate-level: panel of federal election results, 1990-2013; individual-level: SOEP	Within-(former)-GDR comparison of border and non-border areas	Voters who lived close to inner-German border, where life was harder, are *less* likely to lean toward the successor party to East Germany’s communists
Carl (2018): Attitudes toward immigration	SOEP 1999-2016	Duration of life under communism	Concerns about immigration are stronger the longer an individual has spent under communism

(continued)

Table 3 (continued)
Evidence on East-West Differences

Paper, Outcome	*Data*	*Empirical approach*	*Result*
Culture			
Rainer and Siedler (2009): Social and institutional trust	ALLBUS 1991, 1994, 2002	East-West comparison, focus on convergence	East Germans have significantly lower social trust without signs of convergence; institutional trust in East converges towards West
van Hoorn and Maseland (2010): Values	SOEP 1991–2006	East-West differences in transforming situational factors into happiness	East Germans appear to entertain values more conducive to economic growth
Bauernschuster et al. (2012): Self-reliance	ALLBUS 1991, 1994, 1998, 2000, and 2004	East-West comparison	East Germans show lower self-reliance conditional on regional differences in current economic development
Boenisch and Schneider (2013): Preferences for geographic mobility	SOEP 1994	Eastern origin as IV for club membership and church attendance	Those who grew up in the GDR are less likely to be members of clubs or to attend church, which in turn relates to lower geographic mobility
Heineck and Süssmuth (2013): Trust, risk, fairness, cooperation	SOEP 2003 and 2008	East-West comparison, focus on convergence	East-West convergence in risk preferences, less so in trust, close to none in cooperation
Friehe and Mechtel (2014): Conspicuous consumption	Income and expenditure sample 1993, 2008	East-West comparison	Conspicuous consumption more important in East Germany
Möhlmann (2014): Tax morale	World Value Survey 2006; European Values Survey 2008; ALLBUS 2000, 2002; ISSP Religion II 1998; ISSP Citizenship 2004; European Social Survey 2004	East-West comparison, focus on convergence	Persistent gap in tax morale and no sign of quick convergence
Dragone and Ziebarth (2017): Novelty consumption	German National Health Interview and Examination Survey 1991 and 1998	Difference-in-differences: East-West food consumption 1991–1998	East Germans consumed more novel Western food and gained more weight than West Germans when a larger variety of food products became readily accessible after the fall of the Wall
Friehe and Pannenberg (2020): Time preferences	SOEP 2008 and 2013	East-West comparison; treatment intensity; RDD around border	East Germans are less present biased
Gender roles			
Bauernschuster and Rainer (2012): Sex-role attitudes	ALLBUS 1980-2010 (biannual)	East-West comparison, focus on convergence	East Germans are significantly more likely to hold egalitarian sex-role attitudes than West Germans; no evidence of convergence
Fuchs-Schündeln and Masella (2016): Tertiary education	Microcensus, 2005-2008	Difference-in-differences estimation	Additional year of socialist education decreases probability of college degree and affects labor-market outcomes for men
Klüsener and Goldstein (2016): Extramarital fertility	County data 1878, 1937, 2009	East-West comparison over 140 years	Non-marital births the norm in East, but not in West; difference predates 1945 division
Wyrwich (2017): Female labor-force participation	County data 1925, 1939, 1996-2015; ALLBUS 1996-2012	East-West comparison over nearly 100 years	Substantial evidence of persistence in female labor-force participation

(continued)

Table 3 (continued)
Evidence on East-West Differences

Paper, Outcome	*Data*	*Empirical approach*	*Result*
Beblo and Görges (2018): Gender gap in work preferences	ALLBUS 1991, 1998, and 2012	East-West comparison, focus on convergence	Substantial East-West difference in gender gap in work preferences directly after reunification; no convergence thereafter
Campa and Serafinelli (2019): Sex-role attitudes	SOEP and ALLBUS	East-West comparison; RDD around inner-German border	Eastern women more likely to place importance on career success; East Germans less likely to hold traditional gender-role attitudes; effect stronger where female employment growth faster
Lippmann, Georgieff, and Senik (2020): Gender norms	SOEP 1991-2012	Outcomes regressed on dummy for wife higher earner, interacted with East Germany dummy	More equal breadwinner norm in East: women can earn more than their husband without having to increase housework hours, put their marriage at risk, or withdraw from labor market
Lippmann and Senik (2018): Math scores	PISA-E, SOEP	East-West comparison	Gender gap in math achievement is lower in former GDR

Note: Authors' own elaboration based on literature survey. See main text for more detailed discussion.

differences in entrepreneurship seem to be at least partly the result of persistence from before World War II rather than a pure result of communism. Even when taking the broader view that selective migration before 1961 may be regarded as an effect of communism, Fritsch and Wyrwich (2014) document regional persistence of preferences for entrepreneurship between pre-World War II and today. However, not all entrepreneurial spirit is gone. Fuchs-Schündeln (2009) shows that those who are self-employed in the first ten years after the end of communism display higher job satisfaction compared to their retrospective job satisfaction in 1985.

East Germany offered its citizens very limited opportunities to invest. Stocks were the incarnation of a capitalist system that was despised. Laudenbach, Malmendier, and Niessen-Ruenzi (2020) present evidence that East Germans still invest significantly less in the stock market in the 2000s. It seems that the effects are stronger for individuals exposed to stronger communist priming, like those living in communist "showcase cities" or cities of Olympic gold medalists. In contrast, East Germans with negative experiences—those experiencing greater environmental pollution, suppression of religious beliefs, or lack of access to Western television—invest more in the stock market today. These differences are consistent with lasting effects of communism on stock-market participation. There are also effects on investment types: Consistent with communist friends-and-foes propaganda, East Germans are more likely to hold stocks of companies in (ex-) communist countries such as China, Russia, and Vietnam.

Similarly, there is evidence that communism permanently affected savings behavior. Fuchs-Schündeln and Schündeln (2005) show that after reunification, precautionary wealth as a percentage of all wealth in the East is nearly twice that in the West (22.1 versus 12.9 percent). Using the surprise effect of reunification for identification, they exploit differences amongst East Germans with different occupations, based on the idea that the choices made under communism were optimal in that context but may suddenly no longer be optimal under the new environment. Specifically, while labor-income risk under communism was essentially zero, some occupations made for safe civil servant careers after reunification whereas others became risky private-sector jobs. German reunification also constitutes a large unanticipated shock to labor and retirement incomes as well as wealth levels. Fuchs-Schündeln (2008) shows that East Germans have higher savings rates after reunification largely to make up for a perceived gap in retirement savings under the new capitalist regime—consistent with a precautionary saving motive.

Inhabitants of East Germany were used to zero (official) inflation in their centrally planned economy.[7] Reunification came with a fast increase in prices after the abolishment of price controls. Inflation expectations of East Germans continue to be substantially higher than those of West Germans, even decades after reunification (Goldfayn-Frank and Wohlfart forthcoming). Arguably, concerns about preexisting differences in inflation expectations or selective migration on inflation expectations are far-fetched in this case. Instead, these results suggest that the experience of communism and the subsequent transition shock have had long-lasting effects on economic expectations.

Political Preferences

The literature also studies whether life under communism permanently affected political preferences. In their seminal study, Alesina and Fuchs-Schündeln (2007) show that former residents of the German Democratic Republic differ from those of the Federal Republic of Germany in their thinking about market capitalism and the role of the state in providing social services, insurance, and redistribution from the rich to the poor, using data from the German Socioeconomic Panel (SOEP).[8] The authors are careful in considering the identification challenges we described earlier. They point out that before World War II, income per capita did not differ substantially between East and West German states and across Prussian provinces at an aggregate level, to the extent that the regions are separable. They also argue that destruction during World War II was major but universal in both parts of Germany.[9] Addressing inner-German migration, they argue that self-selection motives are unlikely to be strong enough to explain away their effects. Based on our more fine-grained and extensive evidence on pre-World War II differences in communist vote shares and economic structures

[7] Even though state propaganda promoted consumer price stability, residents of East Germany in fact experienced substantial hidden inflation, for example, through changes in product lines. Official price statistics did not include changes from new and enhanced products (Heske 2009, pp. 154ff).

[8] Relatedly, Corneo and Grüner (2002) found that, in 1992, Eastern Europeans had stronger preferences for redistribution than individuals from Western countries.

[9] As discussed above, destruction during World War II was indeed quite similar, but the East was substantially more affected by dismantling and reparations. This might well have affected calls for intervention and redistribution in the early years of East Germany and should arguably be attributed to Soviet occupation as opposed to life under communism in East Germany.

as well as selective migration by occupation and education, it is conceivable that some of the effect attributed to the exposure to the communist political system is the result of preexisting differences and selective migration.

One way to circumvent some of these issues of preexisting differences is to take 1990 as the starting point and focus on *convergence* in attitudes after the fall of communism. Alesina and Fuchs-Schündeln (2007) present evidence of convergence in preferences after reunification, which points to some dynamic influence of political systems in the German context. Similarly, Svallfors (2010) documents considerable convergence in attitudes towards state intervention between East and West Germany using International Social Survey Programme (ISSP) data for the years 1990, 1996, and 2006. But in this case, a flow of Western-socialized people to the East could lead to convergence because of changes in ISSP sample composition over time. Still, the combined evidence leaves little doubt that communism had *some* enduring effect on political preferences. At the same time, it would be unwise to conclude that German reunification can be treated as a natural experiment for *any* outcome of interest.

Corroborating evidence is based on differences *within* East Germany. On some accounts, East Germans living close to the inner-German border tended to have harder lives during the times of German division: authorities implemented forced relocation of whole villages, frequent controls that may have resulted in mental stress, and even more limited freedoms than elsewhere in East Germany. Avdeenko (2018) finds that the successor party to East Germany's communist party captures a lower vote share in the border area than in other parts of East Germany. Concerns about preexisting differences are directly addressed by drawing on 1919–1933 election results. If anything, communist vote shares were higher in the border area than in other parts of the later East Germany, suggesting that the results of a turn away from communism are even stronger once pre-World War II preferences are taken into account. A related methodology uses the number of years under communism as a key variable. The German Socioeconomic Panel includes time-varying information on political attitudes in conjunction with variation in year of birth and hence time under communism and after communism. Carl (2018) shows that those having spent more time under communism in the East, where exposure to foreigners was limited, are more opposed to immigration. Measuring exposure to communism as a continuous variable reduces some of the concerns about preexisting differences and selective migration.

It can also be revealing to compare West Germans in West Germany to West Germans in East Germany and vice versa. Brosig-Koch et al. (2011) run lab experiments with students in the cities of Magdeburg (East Germany) and Essen (West Germany) in 1995 and 2009. East Germans show consistently less fairness and willingness to cooperate in solidarity games, with no convergence in the 20 years after reunification. West German students studying in the East differ significantly from those studying in the West. In fact, they show similarities with East German students. The authors see this as consistent with two potential explanations: West Germans in East Germany either have partly accommodated to the East German behavioral norms or they were more likely to move to the East because they could align themselves better to the social behavior in East Germany—which underlines the importance of considering selective migration.

Overall, papers that exclusively look at the extensive margin of exposure to communism run a risk of over-estimating treatment effects because of pre-World War II differences in communist leanings and selective migration based on preferences for or against communist ideas. However, studies that exploit continuous measures of communist experience (for example, geographic variation or time under communism) and those exploiting convergence in political preferences after the fall of communism suggest rather long-lasting effects on political preferences.

Culture

A large literature looks at cross-German outcomes in the area of culture such as trust, fairness, self-reliance, time preferences, conspicuous consumption, and tax morale. In many areas, cultural values continue to differ between East and West.

Trust is considered to be one of the most important cultural drivers of economic exchange (Arrow 1972). Social trust toward other people is lower in East Germany and does not show signs of convergence. Given lower levels of pre-World War II church attendance as a measure of community interaction, it seems likely that some of this reflects long-term persistence, as opposed to an effect of communism. Evidence from different datasets shows that East Germans are significantly less trusting towards other people than West Germans and that these attitudes are not converging: for discussion, see Rainer and Siedler (2009) for the German General Social Survey and Heineck and Süssmuth (2013) for the German Socioeconomic Panel. Interestingly, however, Rainer and Siedler show that trust in institutions has converged after reunification, suggesting political systems matter in this case. Going beyond trust, Heineck and Süssmuth find that while East Germans are more risk-loving after reunification, there is convergence in risk attitudes between East and West Germans. Looking at differences in perceived fairness and cooperation, they find no evidence of convergence between individuals in the two parts of Germany. To the extent that there is evidence of convergence after reunification in the two studies, concerns about pre-World War II differences and pre-1961 German-German migration are less warranted.

Lower trust in other people can also be seen in higher investment in strong ties such as close friends compared to weak ties such as club membership or church attendance. Using the German Socioeconomic Panel, Boenisch and Schneider (2013) demonstrate that persons who grew up in East Germany exhibit this specific social capital mix, which also corresponds with lower geographic mobility. Given our evidence of lower church attendance in the East well before World War II, some of this presumed effect of communism likely originates in longer-term historic differences.

It is often argued that with its ubiquitous influence on people's lives, communism led to a lack of self-reliance. Bauernschuster et al. (2012) show that this is indeed the case when conditioning on regional differences in economic development. But in line with what we discussed earlier, they also show that the socialist regime affected the composition of the East German population by inducing selective migration before the construction of the Berlin Wall in 1961.

The planned economy of East Germany disappointed its citizens in many ways. As one example, wait times for cars could be well over a decade. Not surprisingly, East German doctrine taught people to live sparingly and strive for improvement of their performance in the future. Consistent with this doctrine, Friehe and Pannenberg (2020) find that former residents of East Germany are less present-biased. While the authors address selective East-West migration of impatient people by using information about parent's region of origin and retrospective information on East-West migration available in a subsample, pre-World War II differences are a potential source of concern. If time preferences vary by some of the dimensions for which we uncovered preexisting East-West differences—say, between blue- and white-collar workers—part of the apparent effect of communism might reflect long-term persistence.

Following the theme of communism-induced limits to consumption choices, some papers exploit the fact that reunification opened up new consumption opportunities. Friehe and Mechtel (2014) show that East Germans display more conspicuous consumption—that is, they spend more on items that display high social status. In a similar vein, Dragone and Ziebarth (2017) show that when a larger variety of food products became readily accessible after the fall of the Wall, East Germans consumed more novel Western food and gained more weight than West Germans. While communism first constrained consumer choice, the transformation period following reunification supported a further differentiation of consumption patterns.

Culture is also reflected in tax morale, that is, the willingness to pay taxes. In fact, tax morale can be seen as a form of solidarity. A stronger preference for redistribution suggested by Alesina and Fuchs-Schündeln (2007) needs the support of honest taxpayers to have the desired consequences. In line with this reasoning, Möhlmann (2014) documents a persistently higher tax morale in East Germany and no sign of quick convergence. We are not aware of data on tax morale before World War II. Yet preferences for communism expressed via higher communist vote shares before World War II and selective migration before 1961 suggest the possibility that East Germans might have had higher tax morale to begin with, in which case the effect of living under communism on tax morale might be overestimated.

Finally, evidence on how life satisfaction responds to circumstances has also been used to measure cultural values. For example, van Hoorn and Maseland (2010) find that getting divorced hurts happiness less for East Germans than it does for West Germans, implying that East and West Germans respond differently to the same circumstance. Considering a wide range of aspects, East Germans appear to be more likely to entertain cultural values conducive to economic performance, such as a stronger dislike of unemployment. The authors conclude that "the belief that economic differences between Eastern and Western Germany are a result of a communist cultural legacy may be largely a myth." We would not go as far, as some of the studies discussed above show strong evidence of enduring effects of communism on cultural values. But our reservation is that the magnitude of effects attributed to the communist regime may in some cases be overestimated.

Gender Roles

Several papers also examine gender-related aspects of the German experience of separation and reunification, including sex-role attitudes, female labor-force participation, gender-specific educational achievement, and family-formation behavior.

East German institutions encouraged female employment, while the West German system deterred women, in particular mothers, from full-time employment. Using the German General Social Survey, Bauernschuster and Rainer (2012) show that East Germans are significantly more likely to hold egalitarian sex-role attitudes than West Germans, with no sign of convergence after reunification. Campa and Serafinelli (2019) present similar results on sex-role attitudes using the German Socioeconomic Panel and a regression discontinuity design focused on areas around the inter-German border. Furthermore, positive attitudes towards female employment are stronger in areas where growth in female employment was larger. Lippmann, Georgieff, and Senik (forthcoming) show that East German women contribute a larger share to household income. At the same time, West German women who contribute more to household income also put in more housework hours.

To what extent might these differences be influenced by pre-separation history of the sort we discussed earlier? Wyrwich (2017) presents detailed analyses of female labor-force participation in 1925 and after World War II and addresses the concern we raised by using difference-in-differences estimation. However, Beblo and Görges (2018) find no substantial gender differences in preferences for work between what would later become East and West in pre-World War II data. The key difference here seems to be that they look at female labor-force participation as a share of the total, whereas it seems more natural to look at female labor-force participation rates, which do differ between East and West (as we show above, following Wyrwich 2019). To the extent that there is substantial persistence in female labor-force participation, at least some of the effect attributed to communism by the literature on gender norms in the labor market may pick up persistence.

Going beyond the labor market, Lippmann and Senik (2018) show that the stereotypical underperformance of girls in math is sharply lower in the regions of the former East Germany, in contrast with those of the former West Germany. The difference is not explained by differences in economic conditions or teaching styles across the former political border. A potential concern about preexisting differences relates to long-lasting Protestant-Catholic differences across Germany and the smaller gender gap in Protestant areas (Becker and Woessmann 2008).

In another study of socialist education, Fuchs-Schündeln and Masella (2016) start from the observation that access to college was restricted under communism. With reunification, high-school graduates had more choice over their post-secondary education. To estimate the effect of socialist schooling on tertiary education, they exploit sharp birth-date cutoffs for school entrance that led to variation in the length of exposure to the East German education system. The authors show that an additional year of socialist schooling decreased the probability of obtaining a college degree for both genders and negatively affected several indicators of the labor-market prospects for men. This methodology is unaffected by the arguments we raised. The results suggest that growing up under socialist education had real effects on labor-market careers, especially for men.

Finally, nonmarital births are the norm in Eastern Germany (58.1 percent of all births in 2009), whereas they are the exception (26.5 percent) in Western Germany (Klüsener and Goldstein 2016). As discussed earlier, differences in nonmarital births predate the 1945 division of Germany, showing that at least some of the East-West difference is likely the result of long-run persistence in attitudes toward gender.

Overall, while the impact of the socialist regime on gender roles seems beyond doubt, numerically some of the results in the literature are likely overestimated because of the preexisting differences in female labor-force participation and the stronger historical drive for gender equality in the largely Protestant East.

Conclusion

The German separation into the communist German Democratic Republic and the capitalist Federal Republic of Germany in 1949 and their reunification in 1990 both happened rapidly and largely unexpectedly. This does not necessarily mean, however, that these events constitute a natural experiment that randomly assigned similar populations to two different political regimes. We show that, in fact, the East and West German populations differed already before World War II. For instance, people in what would later become the communist East were more likely to be working class, less likely to be self-employed, more likely to vote for the communist party, less likely to attend church, and more likely to experience female labor-force participation. To the extent that such differences persist over time, they likely introduce an upward bias in estimated effects of communist exposure on outcomes such as lack of entrepreneurship, left-leaning political preferences, lack of community participation, and equal gender roles.

The East was also more heavily affected by war-related damages, dismantling, and reparation costs to the occupying forces in the immediate aftermath of World War II. Furthermore, roughly one in five people living in the East in 1945 migrated to the West before the construction of the Berlin Wall in 1961. As this out-migration was highly selective, this creates an additional source for potentially persisting East-West differences. Thus, any East-West differences in reunified Germany cannot necessarily be interpreted as a pure effect of communism. Furthermore, evidence of pre-World War II differences persisting over many decades suggests that convergence between the two parts of Germany may take longer than commonly thought.

With this background in mind, is there an effect of communism in the German case? The answer is certainly yes—but each research question requires individual consideration of the aforementioned challenges. More robust evidence for the impact of political systems comes from the convergence of some economic behaviors and political (more than social) attitudes between the two parts of Germany after reunification. Living under the East German regime also seems to have affected consumption patterns persistently. Trust in the state may have been affected by German separation but has converged between East and West after reunification. Gender roles may have been affected along several dimensions, but female labor-force participation and fertility behavior also appear to have a strong component of persistence dating back far before World War II.

The broader lesson is that researchers should not be too quick to take the *formation* of political systems as exogenous. Given that political preferences are endogenous, for instance, to previous experiences (Fuchs-Schündeln and Schündeln 2015), political systems may become endogenous too. This is obvious in the case of revolutions initiated by populations unhappy with their current political system. But even when outside powers initiate a regime transition and when borders are redrawn, any "effect" of the new regime should be carefully assessed with regards to preexisting conditions and selective migration, as those unsatisfied with the regime change might migrate out, leaving behind those who are more disposed to go along with the new system.

■ *Our particular gratitude goes to Oliver Falck, with whom we started this project long ago and who provided continued support throughout. We also thank the four JEP editors—Enrico Moretti, Gordon Hanson, Heidi Williams, and Timothy Taylor—for extensive comments. Discussion and comments from Davide Cantoni, Fredrik Carlsen, Nicola Fuchs-Schündeln, Oded Galor, Ulrich Glogowsky, Peter Gourevitch, Moshe Hazan, Sebastian Klüsener, Kjell Erik Lommerud, Rocco Macchiavello, Omer Moav, Markus Pannenberg, Jared Rubin, André Steiner, Joachim Voth, Romain Wacziarg, Felix Weinhardt, Niko Wolf, Chris Woodruff, Michael Wyrwich, and seminar participants in Trondheim, Bergen, Warwick, and San Diego in 2012, the Warwick Summer Workshop in Economic Growth 2014, and the Workshop on Institutions, Culture, and Long-Run Development 2015 at the Center for Advanced Studies in Munich are gratefully acknowledged. We would like to thank Ulrich Blum, Lucian Hölscher, Sebastian Klüsener, Thomas Rahlf, and Michael Wyrwich for providing data. Christoph Albert and Laurenz Detsch provided capable research assistance. Support by the competitive procedure of the Leibniz Association, the German Science Foundation (CRC TRR 190), and the Fritz Thyssen Foundation is gratefully acknowledged.*

References

Akerlof, George A., Andrew K. Rose, Janet L. Yellen, Helga Hessenius, Rudiger Dornbusch, and Manuel Guitian. 1991. "East Germany in from the Cold: The Economic Aftermath of Currency Union." *Brookings Papers on Economic Activity* 1991 (1): 1–105.

Alesina, Alberto, and Nicola Fuchs-Schündeln. 2007. "Goodbye Lenin (or Not?): The Effect of Communism on People's Preferences." *American Economic Review* 97 (4): 1507–28.

Angrist, Josh. 2002. "How Do Sex Ratios Affect Marriage and Labor Markets? Evidence from America's Second Generation." *Quarterly Journal of Economics* 117 (3): 997–1038.

Arrow, Kenneth J. 1972. "Gifts and Exchanges." *Philosophy and Public Affairs* 1 (4): 343–62.

Avdeenko, Alexandra. 2018. "Long-term Evidence of Retrospective Voting: A Natural Experiment from the German Democratic Republic." *European Economic Review* 103: 83–107.

Baar, Lothar, Rainer Karlsch, and Werner Matschke. 1995. "Kriegsschäden, Demontagen und Reparationen." In Materialien der Enquete-Kommission, Aufarbeitung von Geschichte und Folgen der SED-Diktatur in Deutschland, Vol. 2, 868–988. Munich: Nomos.

Bauernschuster, Stefan, Oliver Falck, Robert Gold, and Stephan Heblich. 2012. "The Shadows of the Socialist Past: Lack of Self-reliance Hinders Entrepreneurship." *European Journal of Political Economy* 28 (4): 485–97.

Bauernschuster, Stefan, and Helmut Rainer. 2012. "Political Regimes and the Family: How Sex-Role Attitudes Continue to Differ in Reunified Germany." *Journal of Population Economics* 25 (1): 5–27.

Beblo, Miriam, and Luise Görges. 2018. "On the Nature of Nurture. The Malleability of Gender Differences in Work Preferences." *Journal of Economic Behavior and Organization* 151: 19–41.

Becker, Sascha O., Markus Nagler, and Ludger Woessmann. 2017. "Education and Religious Participation: City-level Evidence from Germany's Secularization Period 1890–1930." *Journal of Economic Growth* 22 (3): 273–311.

Becker, Sascha O., and Ludger Woessmann. 2008. "Luther and the Girls: Religious Denomination and the Female Education Gap in Nineteenth-Century Prussia." *Scandinavian Journal of Economics* 110 (4): 777–805.

Becker, Sascha O., and Ludger Woessmann. 2009. "Was Weber Wrong? A Human Capital Theory of Protestant Economic History." *Quarterly Journal of Economics* 124 (2): 531–96.

Becker, Sascha O., and Ludger Woessmann. 2013. "Not the Opium of the People: Income and Secularization in a Panel of Prussian Counties." *American Economic Review, Papers and Proceedings* 103 (3): 539–44.

Becker, Sascha O., and Ludger Woessmann. 2018. "Social Cohesion, Religious Beliefs, and the Effect of Protestantism on Suicide." *Review of Economics and Statistics* 100 (3): 377–91.

Boenisch, Peter, and Lutz Schneider. 2013. "The Social Capital Legacy of Communism-Results from the Berlin Wall Experiment." *European Journal of Political Economy* 32: 391–411.

Borjas, George J. 1987. "Self-selection and the Earnings of Immigrants." *American Economic Review* 77 (4): 531–53.

Brosig-Koch, Jeannette, Christoph Helbach, Axel Ockenfels, and Joachim Weimann. 2011. "Still Different after All These Years: Solidarity Behavior in East and West Germany." *Journal of Public Economics* 95 (11–12): 1373–76.

Campa, Pamela, and Michel Serafinelli. 2019. "Politico-Economic Regimes and Attitudes: Female Workers under State Socialism." *Review of Economics and Statistics* 101 (2): 233–48.

Carl, Matthew. 2018. "The Effect of Communism on People's Attitudes toward Immigration." https://ssrn.com/abstract=3246617.

Corneo, Giacomo, and Hans Peter Grüner. 2002. "Individual Preferences for Political Redistribution." *Journal of Public Economics* 83 (1): 83–107.

Dragone, Davide, and Nicolas R. Ziebarth. 2017. "Non-separable Time Preferences, Novelty Consumption and Body Weight: Theory and Evidence from the East German Transition to Capitalism." *Journal of Health Economics* 51: 41–65.

Eder, Christoph, and Martin Halla. 2018. "On the Origin and Composition of the German East-West Population Gap." IZA Discussion Paper 12031.

Fairlie, Robert W., and Magnus Lofstrom. 2015. "Immigration and Entrepreneurship." In *Handbook of the Economics of International Migration*, Vol. 1, edited by Barry R. Chiswick and Paul W. Miller, 877–911. Amsterdam: North Holland.

Falter, Jürgen W. 2014. "Political Cleavages in the Weimar Republic and the Rise of National Socialism." *European Political Science* 13 (1): 106–16.

Falter, Jürgen W., and Dirk Hänisch. 1990. "Election and Social Data of the Districts and Municipalities of the German Empire from 1920 to 1933." GESIS Data Archive, Cologne. ZA8013 Data File Version 1.0.0. https://doi.org/10.4232/1.8013.

Friehe, Tim, and Mario Mechtel. 2014. "Conspicuous Consumption and Political Regimes: Evidence from East and West Germany." *European Economic Review* 67: 62–81.

Friehe, Tim, and Markus Pannenberg. 2020. "Time Preferences and Political Regimes: Evidence from Reunified Germany." *Journal of Population Economics* 33 (1): 349–87.

Fritsch, Michael, and Michael Wyrwich. 2014. "The Long Persistence of Regional Levels of Entrepreneurship: Germany, 1925–2005." *Regional Studies* 48 (6): 955–73.

Fuchs-Schündeln, Nicola. 2008. "The Response of Household Saving to the Large Shock of German Reunification. *American Economic Review* 98 (5): 1798–1828.

Fuchs-Schündeln, Nicola. 2009. "On Preferences for Being Self-employed." *Journal of Economic Behavior and Organization* 71 (2): 162–71.

Fuchs-Schündeln, Nicola, and Paolo Masella. 2016. "Long-lasting Effects of Socialist Education." *Review of Economics and Statistics* 98 (3): 428–41.

Fuchs-Schündeln, Nicola, and Matthias Schündeln. 2005. "Precautionary Savings and Self-selection: Evidence from the German Reunification 'Experiment'." *Quarterly Journal of Economics* 120 (3): 1085-1120.

Fuchs-Schündeln, Nicola, and Matthias Schündeln. 2015. "On the Endogeneity of Political Preferences: Evidence from Individual Experience with Democracy." *Science* 347 (6226): 1145–48.

Goldfayn-Frank, Olga, and Johannes Wohlfart. Forthcoming. "Expectation Formation in a New Environment: Evidence from the German Reunification." *Journal of Monetary Economics.*

Grashoff, Udo. 2019. "Driven into Suicide by the East German Regime? Reflections on the Persistence of a Misleading Perception." *Central European History* 52 (2): 310–32.

Grogger, Jeffrey, and Gordon H. Hanson. 2011. "Income Maximization and the Selection and Sorting of International Migrants." *Journal of Development Economics* 95 (1): 42–57.

Handl, Johann, Karl Ulrich Mayer, Walter Müller. 1975. Mikrozensus-Zusatzerhebung 1971. "Berufliche und soziale Umschichtung der Bevölkerung." Codebook und Grundauszählung für die deutsche Wohnbevölkerung, 15 Jahre und älter. Mannheim, Germany: SPES-Projekt, Universität Mannheim.

Hänisch, Dirk. 1989. "Inhalt und Struktur der Datenbank »Wahl- und Sozialdaten der Kreise und Gemeinden des Deutschen Reiches von 1920 bis 1933«. *Historical Social Research* 14 (1): 39–67.

Heidemeyer, Helge. 1994. *Beiträge zur Geschichte des Parlamentarismus und der politischen Parteien / Kommission für Geschichte des Parlamentarismus und der Politischen Parteien.* Vol. 100, *Flucht und Zuwanderung aus der SBZ/DDR 1945/1949–1961: Die Flüchtlingspolitik der Bundesrepublik Deutschland bis zum Bau der Berliner Mauer.* Düsseldorf: Droste.

Heineck, Guido, and Bernd Süssmuth. 2013. "A Different Look at Lenin's Legacy: Social Capital and Risk Taking in the Two Germanies." *Journal of Comparative Economics* 41 (3): 789–803.

Heske, Gerhard. 2009. "Volkswirtschaftliche Gesamtrechnung DDR 1950–1989: Daten, Methoden, Vergleiche." *Historical Social Research, Supplement* 21: 1–356.

Hölscher, Lucian, ed. 2001. *Datenatlas zur Religiösen Geographie im Protestantischen Deutschland: Von der Mitte des 19. Jahrhunderts bis zum Zweiten Weltkrieg.* 4 Vols. Berlin: Walter de Gruyter.

Jessen, Ralph. 1999. *Akademische Elite und Kommunistische Diktatur: Die Ostdeutsche Hochschullehrerschaft in der Ulbricht-Ära.* Wittingen: Vandenhoeck und Ruprecht.

Klüsener, Sebastian, and Joshua R. Goldstein. 2016. "A Long-Standing Demographic East-West Divide in Germany. *Population, Space and Place* 22 (1): 5–22.

Kowalski, Hans-Günter. 1971. "Die 'European Advisory Commission' als Instrument alliierter Deutschlandplanung 1943–1945." *Vierteljahrshefte für Zeitgeschichte* 19 (3): 261–93.

Laudenbach, Christine, Ulrike Malmendier, and Alexandra Niessen-Ruenzi. 2020. "The Long-lasting Effects of Experiencing Communism on Attitudes towards Financial Markets." NBER Working Paper 26818.

Lippmann, Quentin, Alexandre Georgieff, and Claudia Senik. Forthcoming. "Undoing Gender with Institutions: Lessons from the German Division and Reunification." *The Economic Journal.*

Lippmann, Quentin, and Claudia Senik. 2018. "Math, Girls and Socialism." *Journal of Comparative Economics* 46 (3): 874–88.

Luy, Marc. 2020. Lebenserwartung in Deutschland. Data available via https://lebenserwartung.info.

Data retrieved on February 16, 2020.

Max Planck Institute for Demographic Research (MPIDR) and Chair for Geodesy and Mathematics, University of Rostock (CGG). 2011. "MPIDR Population History GIS Collection." https://censusmosaic.demog.berkeley.edu/data/historical-gis-files. (Partly based on Hubatsch and Klein 1975 ff. and Bundesamt für Kartographie und Geodäsie 2011).

Möhlmann, Axel. 2014. "Persistence or Convergence? The East-West Tax-Morale Gap in Germany." *FinanzArchiv* 70 (1): 3–30.

Müller-Enbergs, Helmut, Jan Wielgohs, Dieter Hoffmann, Andreas Herbst, and Ingrid Kirschey-Feix, eds. 2010. *Wer war wer in der DDR?: Ein Lexikon ostdeutscher Biographien.* Berlin: Ch. Links Verlag.

Parey, Matthias, Jens Ruhose, Fabian Waldinger, Nicolai Netz. 2017. "The Selection of High-Skilled Emigrants." *Review of Economics and Statistics* 99 (5): 776–92.

Rainer, Helmut, and Thomas Siedler. 2009. "Does Democracy Foster Trust?" *Journal of Comparative Economics* 37 (2): 251–69.

Rosés, Joan Ramón, and Nikolaus Wolf, eds. 2018. *The Economic Development of Europe's Regions: A Quantitative History since 1900.* New York: Routledge.

Siebert, Horst. 1991. "German Unification: The Economics of Transition." *Economic Policy* 6 (13): 287–340.

Simpser, Alberto, Dan Slater, and Jason Wittenberg. 2018. "Dead but Not Gone: Contemporary Legacies of Communism, Imperialism, and Authoritarianism." *Annual Review of Political Science* 21 (1): 419–39.

Sleifer, Jaap. 2006. *Planning Ahead and Falling Behind: The East German Economy in Comparison with West Germany 1936-2002.* Jahrbuch für Wirtschaftsgeschichte, Beiheft 8. Berlin: Akademie Verlag.

Steiner, André. 2013. *The Plans that Failed: An Economic History of the GDR.* New York: Berghahn Books.

Svallfors, Stefan. 2010. "Policy Feedback, Generational Replacement, and Attitudes to State Intervention: Eastern and Western Germany, 1990-2006." *European Political Science Review* 2 (1): 119–35.

US Department of State. 1968. Foreign Relations of the United States Diplomatic Papers. Vol. 3, 1945: European Advisory Commission; Austria; Germany. Washington, DC: US Government Printing Office.

van Ark, Bart. 1996. "Convergence and Divergence in the European Periphery: Productivity in Eastern and Southern Europe in Retrospect." In *Quantitative Aspects of Post-War European Economic Growth,* edited by Bart van Ark, Nicholas Crafts. Cambridge: Cambridge University Press: 271–326.

van Hoorn, André, and Robbert Maseland. 2010. "Cultural Differences between East and West Germany after 1991: Communist Values versus Economic Performance?" *Journal of Economic Behavior & Organization* 76 (3): 791–804.

van Melis, Damian. 2006. *"Republikflucht": Flucht und Abwanderung aus der SBZ/DDR 1945 bis 1961.* München: Oldenbourg.

Wagner, Andrea. 2008. *Die Entwicklung des Lebensstandards in Deutschland zwischen 1920 und 1960.* Berlin: Akademie Verlag GmH.

Wolf, Nikolaus. 2009. "Was Germany Ever United? Evidence from Intra- and International Trade, 1885–1933." *Journal of Economic History* 69 (3): 846–81.

Wyrwich, Michael. 2017. "Women and the Labour Market in East and West Germany: Socialist Legacy and Pre-socialist Tradition." Jena Economic Research Paper 2017–015.

Wyrwich, Michael. 2019. "Historical and Current Spatial Differences in Female Labour Force Participation: Evidence from Germany." *Papers in Regional Science* 98 (1): 211–39.

Journal of Economic Perspectives—Volume 34, Number 2—Spring 2020—Pages 172–191

The Long-Term Effects of Communism in Eastern Europe

Nicola Fuchs-Schündeln and Matthias Schündeln

Thirty years after the fall of the Iron Curtain, the countries of Eastern Europe have undergone a massive transformation from centrally planned to market economies and from nondemocratic regimes to democratic ones. The speed of these transformations and the experiences along the way have been very heterogeneous in terms of both initial conditions and outcome variables. As one salient example, consider the transition to membership to the European Union (EU). Already in 2004, just 15 years after the end of communism, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, the Slovak Republic, and Slovenia were welcomed as new members of the EU. Bulgaria and Romania joined three years later, while the last country to become a member of the EU was Croatia in 2013. As of today, Albania, North Macedonia, Montenegro, and Serbia are candidate countries to the EU, Bosnia and Herzegovina has applied for candidate status, and there is an association agreement between Kosovo and the EU.

In this paper, we analyze the long-term effects of communism in Eastern Europe by considering four areas where communist and capitalist doctrines fundamentally differ. We confirm differences in data related to these areas right after the fall of the Iron Curtain and investigate whether these differences still exist (up to)

■ *Nicola Fuchs-Schündeln is Professor of Macroeconomics and Development and Matthias Schündeln is Professor of Development Economics, both at Goethe University Frankfurt, Germany. They are both Research Fellows of the Institute for the Study of Labor (IZA), Bonn, Germany. Fuchs-Schündeln is Research Fellow of the Centre for Economic Policy Research (CEPR), London, UK. Their email addresses are fuchs@wiwi.uni-frankfurt.de and schuendeln@wiwi.uni-frankfurt.de.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.2.172.

30 years after the end of communism in both policies and preferences. Central to communism are strong government interventions in markets, severe limits on political freedom, and low inequality across incomes and genders. Therefore, to analyze the long-term effects of communism, we first document macro indicators related to these areas—that is, broad patterns related to government intervention in markets, political freedom, and inequality—and then analyze preferences pertinent to these defining aspects of communism.[1]

Regarding macro indicators, we first show that in terms of economic freedom, the Eastern European countries started out with lower levels but have now converged nearly to the levels of the West. This holds for a variety of indicators. Second, the same holds true for indices of democratization. Third, gross income inequality in Eastern Europe was lower than in the West in the first years after the fall of the Iron Curtain but rapidly rose to levels comparable to those of the West. At the same time, the social expenditure share in the East lags behind the one in the West, leading to at least similar net income inequality levels in both regions today, if not higher ones in the East. Finally, indicators relating to gender equality in the labor market give a mixed picture: female labor force participation, while initially high in the East, fell to levels below the West, and the gender wage gap is similar in East and West. However, full-time work continues to be the norm for women in the East. In summary, the macro indicators seem to largely indicate rapid convergence of institutions and behavior toward the West.

Do we observe the same convergence for preferences? This question is of interest because the institutional changes in Eastern Europe, like greater market freedom and democracy, will be sustainable only if they have popular support. We focus on the same realms in which communist and capitalist societies differ and for which we document broad convergence in institutions: economic freedom, political freedom, income equality, and gender equality. In all of these areas, we find a lasting impact of communism on preferences.

A developing literature has investigated whether communism can be said to have causal effects on preferences. A large share of these papers focus on East and West Germany, with the notion that German separation and reunification provide a "natural experiment" (for an overview, see Fuchs-Schündeln and Schündeln 2005; Fuchs-Schündeln and Hassan 2016). As discussed in a companion paper in this issue by Becker, Mergele, and Woessmann that focuses on the German experience, one difficult issue here is to separate experience under communism from other long-standing social, economic, and political differences across countries or regions. An analysis based on cohort differences can control for these fixed differences. We offer cohort-based evidence that Eastern Europeans who lived longer under communism differ in their preferences toward governmental involvement in the economy, democracy, redistribution, and female labor force participation in ways that are more in line with the communist doctrine. This finding suggests

[1] For analyses of the transition, see the annual Transition Reports of the European Bank for Reconstruction and Development or Svejnar (2002) for an analysis of the first decade of the transition.

Table 1
Definition of Country Groups

EU East	*Non-EU East*	*West*
Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovak Republic, Slovenia	Albania, Bosnia and Herzegovina, Kosovo, North Macedonia, Montenegro, Serbia	Austria, Belgium, Denmark, Finland, France, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, United Kingdom

long-lasting effects of communism on preferences and is in line with the cohort-based findings by Alesina and Fuchs-Schündeln (2007), Fuchs-Schündeln and Schündeln (2015), and Pop-Eleches and Tucker (2017).

Throughout our discussion, we focus on 17 countries in Central and Eastern Europe, including the countries that formerly belonged to Yugoslavia and the Baltic countries. With the exception of the three Baltic countries Estonia, Latvia, and Lithuania, former Soviet Republics are excluded. For simplicity, we call these countries "Eastern European" or "East." We group them into the two groups shown in Table 1, namely, those countries that today are members of the European Union (the "EU East"), and those that have not yet joined the European Union (the "non-EU East"). The latter group covers the countries of former Yugoslavia, except for Croatia, but plus Albania.

As Western European comparison countries, we focus on the countries belonging to the European Union before the Eastern enlargement (the former "EU-15" countries) but omit Germany, given its history of separation into East and West. The remaining 14 countries are listed in the third column of Table 1 and are simply referred to as "West." Data are not always available for all countries listed in Table 1. The country composition of each of the three country groups for the following analyses is noted separately in each subsection.

A word is in order about the terminology in this paper. Most of the countries that we call former "communist" countries identified themselves, rather, as "socialist" countries as was often reflected in their official names (like the "Czechoslovak Socialist Republic" or "Socialist Federal Republic of Yugoslavia"). Here, we avoid trying to define a line between socialism and communism and simply use "communism" as a broad concept including what these countries referred to as socialism. Characteristics of the communist countries were strong elements of central planning of the economy with a high share of government-owned firms or cooperatives and limited and highly regulated markets. At the same time, communist regimes severely restricted democracy, typically by instituting a de facto one-party regime. The exact extent of communist elements in both the economic and political realms differed significantly from country to country (for discussions, see Kornai 1990; Estrin 1991, 2002). We use "capitalism" as a term defining democratic market economies of the West.

The Long-Term Effects of Communism: Macro Facts

In this section, we document the developments of economic freedom, political freedom, and inequality across incomes and genders in the 30 years after the fall of communism. We use data on the country level from 1990, or the first available year, to 2019, or the most recent available year, thus ideally covering three decades. When building country group averages, we use a consistent set of countries over time in each figure. As a downside, the countries forming the group averages are changing slightly from one variable to another, since not all countries in each group have the full time series available for all variables. We detail the countries underlying the averages in notes and parenthetical comments throughout.[2]

To set the stage, we recap how Eastern European countries have fared in terms of one broad measure of development, GDP per capita, during the last 30 years. The former communist countries have exhibited higher growth rates than the West, but convergence is far from complete. Figure 1 shows average GDP per capita on a log scale in our three groups of countries: the EU East (red), the non-EU East (orange), and the West (blue). The figure shows continuously faster growth for the East than for the West countries over the three decades. For the non-EU East, initial GDP was very low. A number of countries for which we have data from 1990 onwards experienced a deep recession in the early 1990s, but convergence was rapid during the second half of the 1990s. The initial GDP level for the EU East was substantially higher.

However, convergence is far from complete: on average, GDP per capita in the EU East is 61 percent and in the non-EU East, 29 percent of that in the West in 2018. For context, Greece, the poorest country in the West, has a GDP per capita between those of Romania and Latvia. The Czech Republic, the richest country in the East, has a GDP per capita between those of Portugal and Spain. GDP per person employed, a rough measure of labor productivity, in the EU East is 58 percent of that in the West, and in the non-EU East, 36 percent of that of the West in 2018.

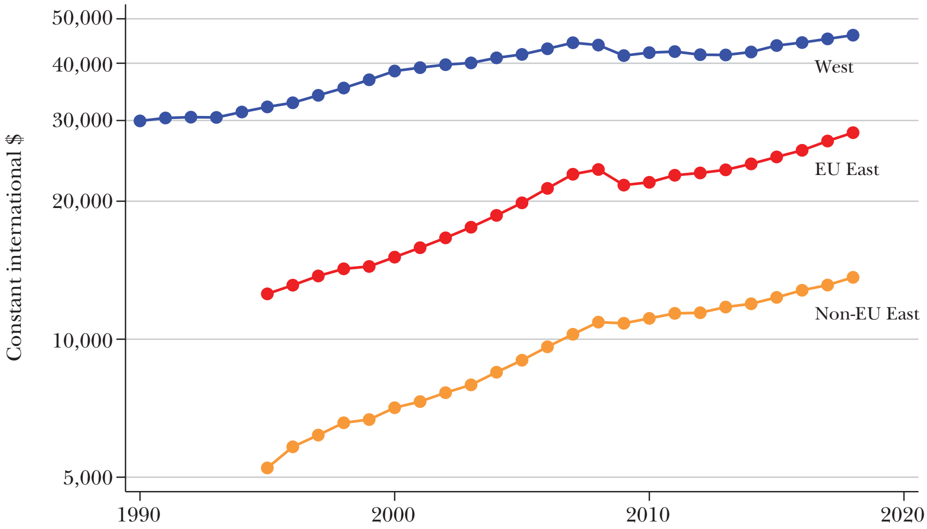
Economic Freedom

Communist regimes impose economic structures based on central planning and a lack of free markets. On measures of economic freedom, the Eastern economies have evolved substantially toward Western Europe with only small gaps remaining.

Panel A of Figure 2 shows the development of the Economic Freedom index of the Heritage Foundation, which is the summary of four subscores that cover four areas: rule of law, government size, regulatory efficiency, and open markets. It ranges from 0 to 100, with higher scores indicating greater economic freedom. As the figure shows, economic freedom was still lower in the EU East than in the West

[2] We always make sure not to omit a large part of the group population by excluding countries. Specifically, we do not show any non-EU East graphs without Serbia, which accounts for roughly 40 percent of the population in the non-EU East group, nor any EU East graphs without Poland, which accounts for roughly 35 percent of the population in the EU East group. Our comparisons all focus on non-population-weighted country-group averages, but we have confirmed that population-weighted averages look largely similar.

Figure 1
GDP per Capita
(*logarithmic scale*)

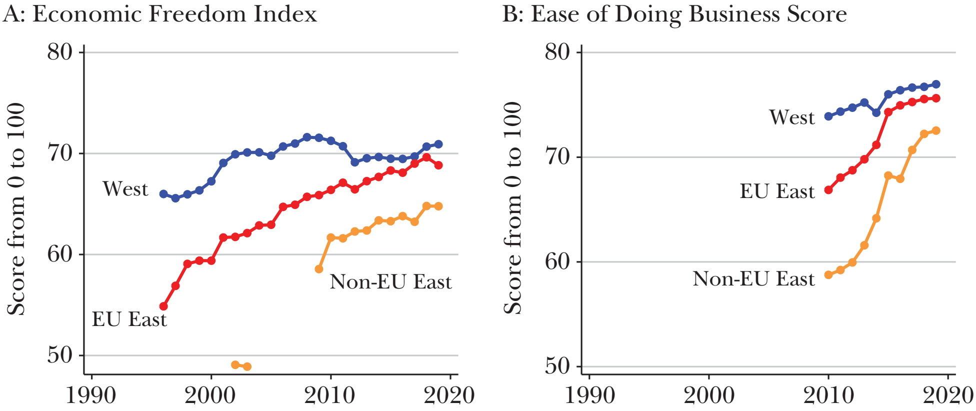


Source: World Bank.
Note: Kosovo and Montenegro are omitted from the non-EU East. The figure has a log scale and shows GDP per capita in constant international 2011 dollars, purchasing power parity adjusted.

in the mid-1990s. In the early 2000s, when data are available also for the non-EU East countries, the non-EU East has the lowest scores of the three groups. However, the East experienced a substantial catch-up in economic freedom over the entire time period, continuing up until the present. In 2019, the EU East exhibits average scores close to the West, and the gap between the non-EU East and the West has also closed considerably. A similar picture emerges from the Ease of Doing Business score provided by the World Bank, which summarizes information on the ease of starting a business, dealing with construction permits, getting electricity, getting credit, and so on, as shown in Panel B of Figure 2. Communist countries largely relied on state-owned enterprises and suppressed the formation of private businesses, so one would expect an initially low score of this summary index in the East after the fall of the Iron Curtain. The index is unfortunately available only from 2010 onwards, but shows initial large differences as well as substantial but incomplete convergence among the three country groups, similar to the index of the Heritage Foundation.[3]

[3] Only two of the 17 East countries have an Ease of Doing Business score lower than Greece—the country with the lowest score in the West. Only six East countries score lower than Italy—the country with the second-lowest score in the West. North Macedonia, the country with the highest score in the East, is surpassed only by the United Kingdom and Denmark in the West.

Figure 2
Indices of Market Orientation



Source: Panel A: Heritage Foundation (https://www.heritage.org/index/); panel B: World Bank.
Note: In the left panel, Kosovo is missing from the non-EU East; Montenegro and Serbia have missing values from 2004 to 2008, thus these years are omitted.

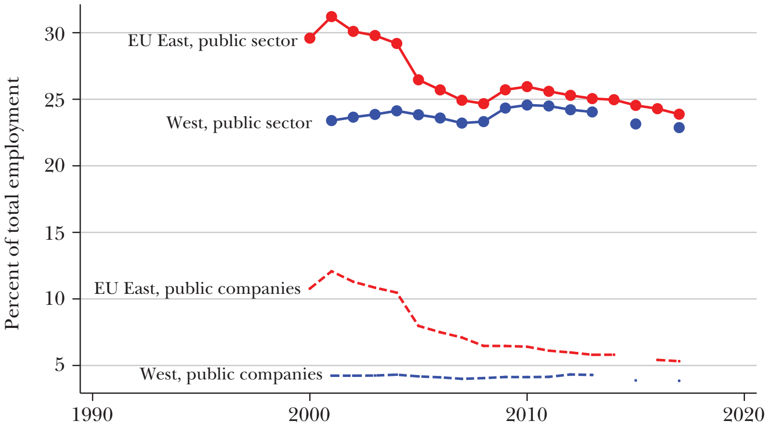
Strong government intervention into markets implies a large number of employees working in the public sector. Figure 3 shows the share of all employees working in the public sector as well as the share working specifically in public companies. The former adds the share of workers in general government to the latter. These data come from the International Labour Organization and are missing for the non-EU East. As Figure 3 shows, the share of workers in public companies and the share of workers in the public sector in total were both substantially higher in the East than in the West in the early 2000s but have been falling steadily in the East toward levels similar to the West. Thus, public employment shows convergence of East levels to the West, mimicking the ones of the indicators of economic freedom.

Democratization

Turning from the economic to the political sphere, we use scores from the Polity IV Project provided by the Center for Systemic Peace, which are a commonly used indicator for the quality of democratic institutions. This index ranges from −10 to 10, with −10 representing hereditary monarchies and 10 consolidated democracies. It is based on different components measuring the quality of executive recruitment, the constraints on executive authority, and political competition. As Figure 4 shows, the EU West countries exhibit an average score above 9.7 throughout the last three decades. The initial score for the EU East in 1993 is below 8, and for the non-EU East it is below 2. By 2006, all three country groups exhibit average scores of 8.8 or above. Thus, according to these indicators, all three country groups feature fairly solid democratic institutions today.

Figure 3

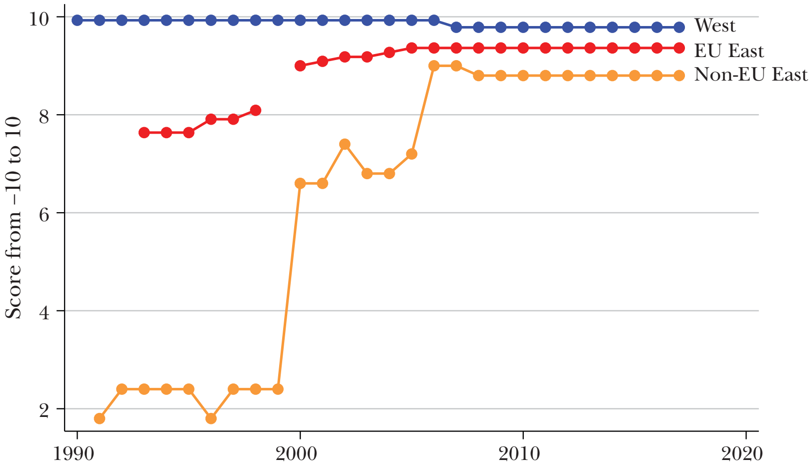
Public Sector Employment



Source: International Labour Organization.
Note: Bulgaria, Croatia, Hungary, and Romania are missing from the EU East. The West comprises Denmark, Greece, Ireland, Luxembourg, Spain, Sweden, and the United Kingdom. Data for Austria, Belgium, France, the Netherlands, and Portugal are not available for the full time-series but line up well with the West average for the available years. Data for 2014 and 2016 are missing for Denmark as well as data on public company employment in 2015 for Latvia, thus these years are omitted from the West and EU East, respectively.

Figure 4

Index of Democratization (Polity Score)



Source: Center for Systemic Peace (http://www.systemicpeace.org/polityproject.html).
Note: Data on Bosnia and Herzegovina are missing. The Center for Systemic Peace provides an aggregate score for Yugoslavia from 1990 to 2002, which we assign to Kosovo, Serbia, and Montenegro. From 2003 to 2005, only a combined score is provided for Serbia and Montenegro, which we assign to both countries and Kosovo. From 2006, individual country scores are available for Serbia and Montenegro, and we assign the score for Serbia to Kosovo for 2006 and 2007. Data on North Macedonia are available throughout as well as data on Kosovo from its foundation in 2008 onwards. The value for Croatia in 1999 is missing, therefore we omit 1999 for the EU East.

Income Inequality and the Welfare State

Equality in economic conditions is a stated goal of communism. The data suggest that pre-tax inequality was indeed lower in the East than in the West at the start of the transition but then increased rapidly. Moreover, government redistribution programs through taxes and transfers have not reached the scale in the East that they have in the West despite the increasing pre-tax inequality. As a result, in terms of post-tax inequality, the East seems to have roughly reached the West's levels. An important caveat for this topic is that measures of inequality are scarce. Future research using different measures to quantify inequality trends would be useful.

To document income inequality, we look at the pre-tax income share belonging to the top 10 percent income earners, using data from the World Inequality Database.[4] In 1990, the share of pre-tax income going to the top 10 percent was slightly above 27 percent in the West, but less than 23 percent in the EU East excluding Poland (for which data are only available as of 1992), and slightly above 23 percent in the non-EU East. However, the share of income of the top 10 percent earners in the EU East caught up with the West at levels of around 29 percent by the mid-2000s and rose to levels above 28 percent in the non-EU East by then.

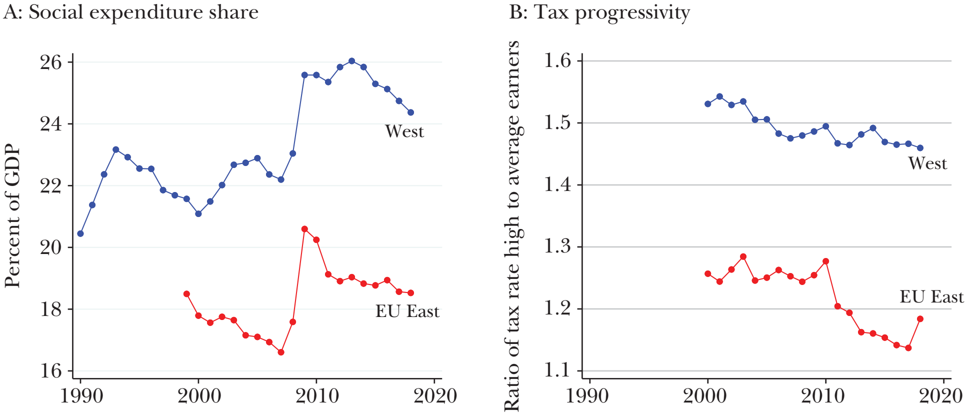
For net inequality, redistribution through taxes and transfers matters. Panel A of Figure 5 shows the average share of social expenditure relative to GDP, using OECD data. Unfortunately, these data are not available for the non-EU East countries. The EU East share always was and continues to be substantially below the share in the West.[5] Panel B of Figure 5 displays a measure of tax progressivity, namely, the ratio of the average income tax rate (combining central and local taxes) at 167 percent of average earnings to the tax rate at average earnings (for a single person without children). While information for the non-EU East is again not available, this ratio is substantially higher in the West than in the EU East, with the difference even increasing over time. A similar picture emerges if we calculate progressivity of the total tax wedge, which also includes social security contributions by both employees and employers.

It might seem somewhat surprising that redistribution through taxes and transfers is apparently lower in Eastern European countries than in Western ones, given the goal of equality under communism. This could, however, be the case because during communism, one important way in which equality was achieved was via direct wage and price regulation, generating less need for classical means of government redistribution in unregulated markets like progressive taxation and social transfer programs. Indeed, the only two countries with data from 1990—Poland and the Czech Republic—both exhibited a share of social expenditure to GDP of 14 percent

[4] The World Inequality Database is at https://wid.world/. Albania, Bosnia and Herzegovina, and Kosovo are missing from the non-EU East comparison.

[5] OECD data also show that the net replacement rate during unemployment was lower in the EU East (specifically in the Czech Republic, Hungary, Poland, and the Slovak Republic, which are the countries with available data) than in the West in the early 2000s, the earliest years with some information, only catching up in the late 2000s.

Figure 5
Measures of the Welfare State and Redistribution



Source: OECD.
Note: Bulgaria, Croatia, and Romania are missing from the EU East in both panels. Panel B shows the ratio of the average income tax rate (combining central and local taxes) at 167 percent of average earnings to the tax rate at average earnings for a single person without children.

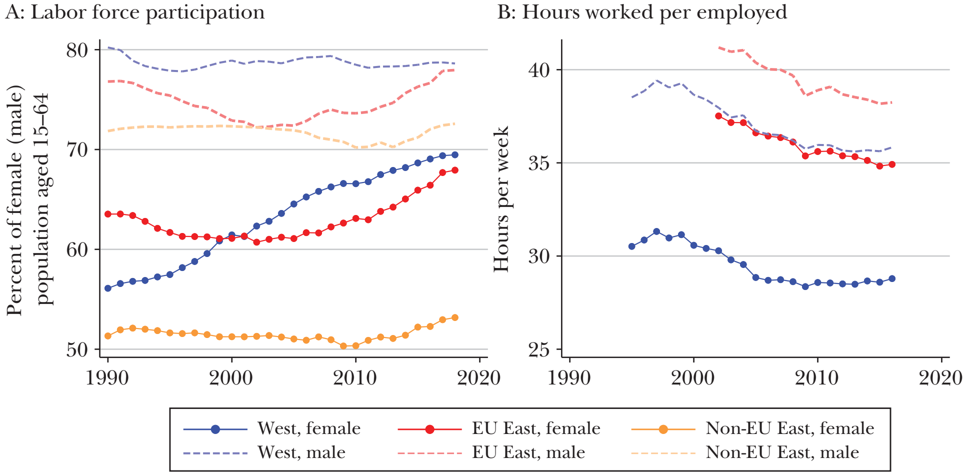
in that year, substantially lower than the current one. At the same time, corporate tax rates are also low in the East, and international tax competition might be another reason for low redistribution (for example, Cassette and Paty 2008).

The evidence of only slightly lower pre-tax inequality in the East, combined with less redistributive measures in the East, would suggest similar or even higher post-tax inequality in the East than in the West. The EU Statistics of Income and Living Conditions (EU-SILC) provide estimates of post-tax income inequality based on household surveys and can be used to analyze this issue. The EU-SILC start in 2004 with a limited set of Eastern European countries. Overall, the average Gini index of post-tax household disposable income in 2017 amounts to 30 in the West, 31 in the EU East, and 35 in Serbia and Macedonia (the only two non-EU East countries for which this statistic is available). Pensions play a significant role here: omitting pensions from disposable income, the Gini coefficient is lowest in the EU East with 34, followed by 36 in the West, and 39 in the non-EU East. Thus, disposable income inequality seems indeed to be at least at a similar level nowadays in the East and in the West, if not higher in the East.

Gender Equality in the Labor Market

A high rate of female labor force participation is a specific and salient feature of communism. The participation of women, especially mothers, in the labor market was actively advocated during communism (Campa and Serafinelli 2019; Fuchs-Schündeln and Masella 2016). Panel A of Figure 6 shows the female labor force participation rate among the population aged 15–64. It was 8 percentage points higher in the EU East than in the West in 1990, while it was always lowest in the

Figure 6
Labor Force Participation and Hours Worked per Employed by Gender



Source: Panel A: The World Bank; Panel B: authors' calculations based on EU-LFS.
Note: Data refer to the population aged 15 to 64. Kosovo is missing from the non-EU East.

non-EU East. However, the female labor force participation rate in the EU East fell during the 1990s, while it rose in the West. This drop in female labor force participation in the EU East during the early transition years was plausibly driven at least partly by the labor demand rather than by the labor supply side: it is smaller than the corresponding drop in the male participation rate (depicted as a dotted line), which in fact fell by 4.6 percentage points (compared with 2.8 percentage points for women) and also reached its trough in 2002. Only toward the mid-2000s did the female labor force participation rate start increasing again in the EU East, and in 2010, it finally started increasing slightly in the non-EU East. Today, labor force participation of men in the EU East lies 0.7 percentage points below the West, while the gap for women is 1.5 percentage points.

The labor force participation rate alone gives an incomplete picture of the labor supply behavior because it blurs any distinction between part-time and full-time work. Throughout the last three decades, women in the EU East have been much less likely to work part-time than women in the West: the OECD reports an average female part-time employment rate of between 23 and 29 percent during the last decades in the West, but only 7 to 9 percent in the EU East (for the years 2002–2018). Based on the European Labor Force Survey (EU-LFS), we can calculate hours worked per employed woman aged 15 to 64, starting in 2002 for the EU East. While hours worked per employed woman have been falling in both the EU East and the West from 2002 onwards, the difference between both regions has been rather constant at seven more hours per week in the East than in the West. The corresponding difference for employed men is only three hours, as Panel B of Figure 6 shows. In this area, we do not observe any convergence between East and West. As a result, total hours worked per woman aged 15 to 64, combining the employment rate and hours worked per employed woman, continue to be significantly higher in the East than in the West, with a difference of around 3.1 hours for the years 2002–2016 (see also Bick, Brüggemann, and Fuchs-Schündeln 2019; and Bick and Fuchs-Schündeln 2018 for further evidence on female employment, part-time work, and part-time regulation in Europe).

Our last measure of gender equality in the labor market is the gender pay gap. On average, male hourly wages in industry, construction, and services (excluding public administration and defense) are 14 percent higher than female ones in 2017. This gender pay gap is almost identical in the EU East and the West according to Eurostat data. However, we lack data for the early transition years, and the caveat applies that this is an unadjusted gender pay gap.

Thus, in the area of gender equality, the picture that different indicators offer is very mixed: the female labor force participation rate points toward not only convergence but even a reversal of East and West. The gender pay gap is nowadays the same in the East and the West, but hours worked per employed woman are constantly higher in the East than in the West.

The Long-Term Effects of Communism: Preferences

Summarizing the findings so far, we observe substantial convergence of the East toward the West in terms of macro indicators. However, institutional changes in a democratic society can last only if they have broad public support. Thus, we now turn from macro indicators to measures of individual preferences. Did communism lead to differences between East and West in attitudes toward the market economy, political institutions, income inequality, and gender equality in the labor market? And do these potential differences still exist even 30 years after the end of communism? Have these attitudes converged as well? We document average differences in preferences between former communist countries and countries from Western Europe. These differences are in line with differences between the communist and capitalist doctrines, except in the case of gender equality in the labor market. However, they could be indicative of both an effect of communism or of preexisting structural or institutional differences between the country groups (due to differences in culture, precommunist history or institutions, or geography, as discussed by the companion paper by Becker, Mergele, and Woessmann). We therefore focus on cohort comparisons in which we exploit variation in exposure to communism across different cohorts within East and West countries. We use different datasets, as detailed below, and seek to use the most recently available data for each dataset.

Here, we rely on a nontechnical argument, using graphical analyses. In the online Appendix available with this paper at the *Journal of Economic Perspectives* website, we make our argument in a more rigorous form with a regression analysis. We show that the cohort-based results presented below hold up in a regression framework that includes all available years and countries of each dataset, adds country-year fixed effects, and controls for individual-level characteristics, in particular, gender and current unemployment status.

Attitudes toward the Market Economy

To study attitudes toward markets, we use data from the Life in Transition Surveys (LITS) 2010 and 2016, which were collected by the European Bank for Reconstruction and Development in collaboration with the World Bank. LITS cover all Eastern European countries and a few Western European comparison countries.[6] We add the 2010 survey round to the 2016 one to expand the set of Western comparison countries. The 2016 round covers only Greece and Italy while the 2010 round covers France, Italy, Sweden, and the United Kingdom.

Specifically, we use the LITS question, "With which one of the following statements do you agree most?" We code a variable, "support for market economy," which is 1 if the respondent chooses "A market economy is preferable to any other form of economic system" and 0 if the respondent chooses "Under some circumstances, a planned economy may be preferable to a market economy" or "For people like me, it does not matter whether the economic system is organized as a market economy or as a planned economy." We code this variable as missing if the respondent chooses "Don't know."

Panel A of Figure 7 shows the percentage share of respondents expressing preferences for the market economy. As before, EU East averages are shown in red, non-EU East averages in orange, and West averages in blue.[7] In contrast to the previous figures, the x-axis in Figure 7 does not depict a timeline, but rather, splits the sample of individual respondents into four cohort groups: those born before 1945, between 1945 and 1959, between 1960 and 1974, and 1975 and after.

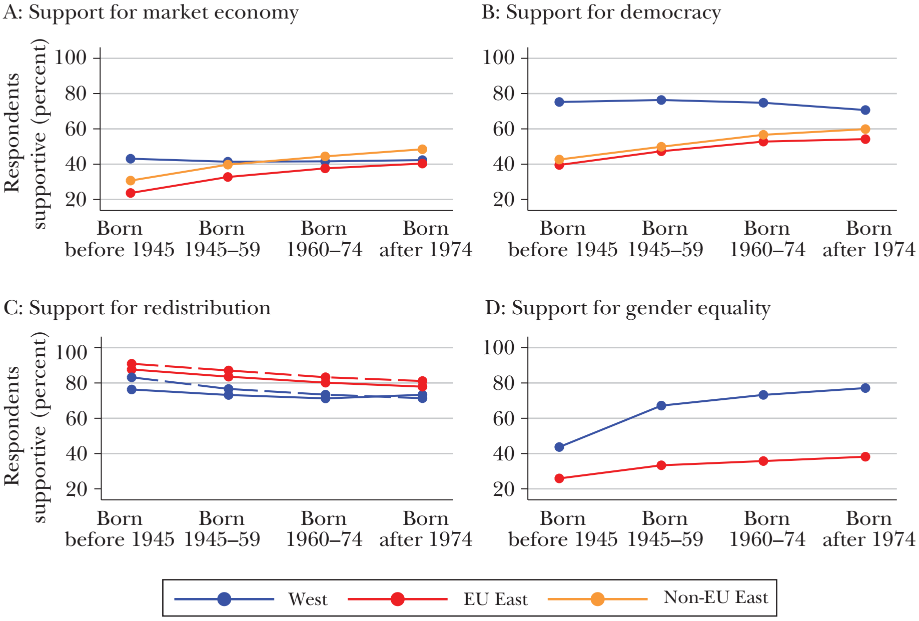
In all three country groups, less than 50 percent of the population agree with the statement that a market economy is preferable to any other form of economic system. On average, respondents in the included West countries express somewhat higher preferences for a market economy than respondents in EU East countries, but very similar preferences to the ones of respondents in non-EU East countries. On the one hand, it is surprising that the support for the market economy is generally so low, given the high degree of market freedom in all three country groups. On the other hand, the similarity of the expressed average preferences in East and West is quite in line with the almost complete convergence in the indices on market freedom we document above. In results not presented here (but available in the online Appendix), considering trends in the average preferences over the years 2010 and 2016, we do not observe any convergence, but rather, stable differences between the EU East and the West, and their stable absence between the non-EU East and the West.

[6] More detail and access to the LITS data can be found here: https://www.ebrd.com/what-we-do/economic-research-and-data/data/lits.html.

[7] To calculate averages for these three groups, we use survey weights to construct country averages and then give each country the same weight like we do when building the macro averages.

Figure 7

Cohort Preferences: Support for Market Economy, Democracy, Redistribution, and Gender Equality



Source: Authors' calculations, based on the following data sources: Panels A and B: Life in Transition Surveys 2010 and 2016; Panel C: International Social Survey Programme, Social Inequality module from 2009 (solid line) and Role of Government module from 2016 (dashed line); Panel D: International Social Survey Programme, Family and Changing Gender Roles module from 2012. The set of countries covered by the data differs for each panel. Details are mentioned in the text.

Any East-West difference in preferences could have existed, however, before communism, rather than being an effect of living under different systems for up to 50 years. One way to get some insight into causality is to look at preferences of different cohorts. The central idea is that individuals may acquire preferences over time through experience. This endogeneity of preferences has been demonstrated for economic preferences by Alesina and Fuchs-Schündeln (2007) and Malmendier and Nagel (2011) and for political preferences by Fuchs-Schündeln and Schündeln (2015). Thus, if communism causally leads to differences in preferences, those who lived for a longer period under communism in a certain country will differ in their preferences from those who experienced communism for a shorter time. In particular, those who lived longer under communism will have preferences that are more aligned with the realities experienced under communism.

We therefore turn to differences across cohorts. Regarding support for the market economy, panel A of Figure 7 shows that there is almost no difference across cohorts in the West: 41 to 43 percent of the respondents are promarket on average for all four cohort groups. On the other hand, in both EU and non-EU East countries, there is a clear cohort gradient. The oldest cohorts (which have longer experience with communism) are much less promarket than the younger cohorts (which have lived fewer years under communism). The difference between these two cohorts in both East country groups amounts to around 17 percentage points. This strong difference in preferences over cohorts is consistent with the hypothesis that communism causally leads to differences in preferences. Given that the youngest cohorts have preferences similar to the ones in the West, or even stronger preferences for a market economy in the non-EU East, these differences in attitudes, however, can be expected to vanish over time.

Attitudes toward Democracy

In the area of support of the market economy, we observe an effect of communism on preferences expressed through a strong cohort gradient in East countries, but also similar preferences between East and West for the youngest cohort group. Do we observe similar facts for the support for democracy? For attitudes toward a democratic system, we use a LITS question similar to the one concerning attitudes toward the market economy. The question asks, "With which one of the following statements do you agree most?" We code the variable "support for democracy" as 1 if the respondent chooses "Democracy is preferable to any other form of political system." The variable is coded 0 if the respondent chooses one of the two other options, namely "Under some circumstances, an authoritarian government may be preferable to a democratic one" or "For people like me, it does not matter whether a government is democratic or authoritarian." If the respondent chooses "Don't know," the variable is coded as missing.

Differences in support for democracy are much more striking between East and West than differences in support for the market economy. As Panel B of Figure 7 shows, in the available West countries (France, Greece, Italy, Sweden, and the United Kingdom), on average 74 percent of the respondents agree with the statement that democracy is preferable to any other form of political system, while only 50 percent of the respondents in the EU East and 56 percent in the non-EU East agree with the statement. Therefore, despite the rapid convergence of the political systems, support for democracy is much weaker in the East than in the West. As is the case for support for the market economy, the support for democracy is higher in the non-EU East than in the EU East countries but remains substantially below the support in the West. The online Appendix provides an analysis of average preferences over time. There, we observe slight convergence of preferences between East and West from 2010 to 2016. The support for democracy of only 74 percent in the West may be surprisingly low. We note that it is in line with previous findings in the literature based on other data sources like the World Values Surveys (for example, see Foa and Mounk 2016).

The cohort patterns regarding support for democracy are very similar to the ones regarding support for the market economy and, again, indicate a lasting effect of communism. There is little variation across cohorts in the West, yet a fairly strong increase in support for democracy across cohorts in the East. The support for democracy in both EU East and non-EU East is 15 to 17 percentage points lower among the oldest cohort group (born before 1945) than among the youngest cohort group (born in 1975 or later). Again, these results suggest a causal effect of communism on preferences for a specific (nondemocratic) political system, which still affects individuals up to 26 years after the end of communism. Moreover, even the youngest cohort groups, having lived at most 15 years under communism, show substantially lower support for democracy in the East than in the West. Therefore, unlike with preferences for the market economy, it is less clear that average differences between East and West will fade out over the coming decades.

Attitudes toward Inequality and Social Policies

We documented a rapid increase in inequality in the East as well as less redistributive policies in the East than in the West. A desire to achieve equality is near the core of the communist ideology. Therefore, if preferences are affected by the system that individuals experience, and this effect is long-lasting, we expect that individuals in former communist countries prefer government policies that promote economic equality, which would be in contrast to the observed policies.

To study preferences related to inequality and social policies, we resort to data from the International Social Survey Programme (ISSP).[8] Two different ISSP modules include questions related to government policies regarding income inequality. The ISSP Social Inequality module from 2009, which we use as a baseline and show in solid lines in Panel C of Figure 7, covers all EU East countries but Romania, and all West countries but Greece, Ireland, Luxembourg, and the Netherlands. Unfortunately, ISSP does not cover any non-EU East countries. We use the question "To what extent do you agree or disagree with the statement, 'It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes'?" Answer categories are "strongly agree," "agree," "neither agree nor disagree," "disagree," and "strongly disagree." We code respondents as "pro redistribution" if they answer either "strongly agree" or "agree" to the question. An alternative measure comes from the ISSP Role of Government module from 2016, namely, the question, "On the whole, do you think it should or should not be the government's responsibility to reduce income differences between the rich and the poor?" This module covers fewer countries and specifically leaves out Poland, the largest EU East country, as well as Bulgaria, Estonia, and Romania. Among the West, the module does not cover Austria, Greece, Ireland, Italy, Luxembourg, the Netherlands, and Portugal. Again, no data on non-EU East countries are available. Results based on this module are shown in dashed lines.

[8] The ISSP data are collected by a cross-national collaboration of academic organizations, universities, and survey agencies and have been available annually since 1985 with rotating topics. We use the modules "Social Inequality" from 2009, "Role of Government" from 2016, and "Family and Changing Gender Roles" from 2012. More detail and access to the ISSP data can be found at *http:// issp.org*. We do not use LITS for this part of the analysis because the available LITS questions do not address general preferences but refer to actions related to the specific home country situation in a given year. Therefore, the answers are less comparable across countries. Relating to gender equality, LITS does not have questions comparable to the ISSP.

Panel C of Figure 7 first shows that across Europe, a large majority of respondents favor redistribution. Comparing East and West, the panel also shows that a larger share of Eastern Europeans are in favor of redistribution than Western Europeans. Relating to the baseline results from the Social Inequality module, on average, the difference between East and West amounts to 8 percentage points. This is in contrast to actual redistributive policies, which are weaker in the East than in the West. Thus, differences between East and West in policies and preferences do not line up in this area. We can also observe answers from the same module for the years 1992 and 1999, and there is no indication of convergence in preferences (as documented further in the online Appendix).

But is the high support for redistribution in the East an effect of communism, or are the East-West differences caused by other reasons? Again, we turn to cohort patterns in order to get an insight into causality. As in Panels A and B, Panel C shows in the baseline results almost no cohort pattern in support for redistribution in the West: around 73 percent of respondents of all cohort groups in the available Western European countries express support for redistribution. By contrast, there is a cohort pattern in the East: 88 percent of the oldest cohort group express their support for redistribution, but only 78 percent of the youngest cohort group. Again, the longer a cohort lived under communism, the more it sees redistribution as an important task of the government, in line with an effect of the communist doctrine on preferences of those who lived longer in a regime that followed this doctrine. For the youngest cohort group, we still observe East-West differences, but they are relatively small.

However, the differences between the cohort gradients in East and West are substantially smaller in the question relating to support for redistribution shown in Panel C than in the questions relating to support for the market economy or support for democracy shown in panels A and B. Also, in the answers to the alternative question from the Role of Government module (shown in dashed lines) the East and West cohort gradients are almost identical. Thus, the evidence regarding long-term effects of communism is weaker in this area than in the previous ones. One reason might be that generally support for redistribution is very high in both East and West.

Attitudes toward Gender Equality in the Labor Market

Finally, we turn to attitudes toward gender equality in the labor market. Communist regimes actively promoted female labor force participation. Did this lead to persistently more favorable attitudes toward working women? We use data from the ISSP survey on the Family and Changing Gender Roles module from 2012 to investigate this. These data cover all EU East countries but Estonia and Romania, and all West countries but Greece, Italy, and Luxembourg.[9] The ISSP survey asks for agreement to the statement, "A man's job is to earn money; a woman's job is to look

[9] We additionally omit data from Spain, which features different answer categories in our relevant question than the rest of the countries.

after the home and family." We code answers of "strongly disagree" or "disagree" as favoring gender equality.[10]

Results for the comparison of East versus West countries are displayed in Panel D of Figure 7. As this panel shows, individuals in the West have on average a substantially more positive attitude toward participation of women in the labor market than in the East. Of all four preference measures, this one shows the largest East-West difference, and it is the only one not showing average preferences in line with communism versus capitalism. On average, in the West around 70 percent of the population support gender equality in the labor market according to our measure, but in the East only 35 percent do. Looking at older data, support for gender equality was higher in the West than in the East back in 1994, and there has been little convergence of preferences between Eastern and Western Europeans for the years for which we have data available (1994, 2002, and 2012, presented in the online Appendix). These findings on preferences are in line with the rapid decrease in female labor force participation in the East in the 1990s, and the higher female labor force participation rate in the West than in the East nowadays. However, they are surprising in light of recent literature, which has established substantially higher female labor force participation and more positive attitudes toward working women in East than in West Germany, and has linked these causally to communism (Campa and Serafinelli 2019; Lippmann, Georgieff, and Senik forthcoming). Indeed, if we analyze the answers to the ISSP statement for Germany, we find higher support for gender equality in the East than in the West (as shown in the online Appendix).

One possible explanation for this divergence of results is that a comparison of averages does not, of course, establish causality. There are many differences between countries in the West and former communist countries in the East that may explain differences in attitudes, and communism is just one of these differences. For example, it could be the case that residents of Western countries have always been more favorable toward gender equality than residents from Eastern countries, and that such differences predate the establishment of communism. Religious differences might be one plausible preexisting factor that could generate differences regarding attitudes toward female labor force participation; in this context, it is intriguing that Djankov, Nikolova, and Zilinsky (2018) find that religious denomination matters in the context of happiness during transition. Research using an East-West comparison within Germany is a way to minimize some of these potential differences.

Here, we turn again to an analysis of cohort differences. In contrast to the other three panels in Figure 7, it is noteworthy that Panel D shows a strong cohort effect in the West when it comes to preferences for gender equality: almost 80 percent

[10] In the online Appendix, we also build an index of gender equality that combines the level of agreement of respondents with seven statements relating to gender equality in the labor market. In addition to the question we use here, these include statements such as "A working mother can establish just as warm and secure a relationship with her children as a mother who does not work" or "Both the man and woman should contribute to the household income." Our findings based on the index are very similar to the findings presented here.

of the youngest cohort group in the West express support for gender equality in our measure, but only 45 percent of the oldest cohort group. One might expect a cohort gradient in preferences toward gender equality, but the difference is very large. In fact, a cohort gradient is also present in the East, but it is much weaker there: the difference in preferences for gender equality is 12 percentage points between the oldest and youngest cohort in the East, but 33 percentage points in the West. This is in line with a lasting effect of communism on preferences toward gender equality: having lived under communism tilted the preferences of older Eastern Europeans toward a preference for gender equality, so the cohort gradient in the East is weaker than in the West. Thus, we conclude that despite the substantially lower average support toward working women in the East than in the West, communism might have had a lasting impact. The differences in average support could be due to preexisting preferences or fundamental differences in the labor market after the fall of the Iron Curtain.

In summary, the cohort analyses suggest a lasting causal impact of communism on support for the market economy, democracy, redistribution, and gender equality.[11]

Conclusion

Communism implies strong government intervention in markets, severe limits on political freedom, and low inequality across incomes and genders. In this paper, we find a fairly rapid convergence of the East to the West for most variables related to these areas: indices for market freedom and the state of democracy, as well as measures of inequality. Today, the differences between East and West in these variables are small or even nonexistent. Regarding gender equality in the labor market, we find that female labor force participation in the EU East rapidly dropped to levels even below the West, although full-time work for those women who are in the labor force continues to be the norm in the East.

Turning from macro indicators to preferences, we show that differences in preferences between East and West on key economic and policy issues persist a quarter century after the end of communism. Older cohorts in the East, who have lived under communism for a longer time, show preferences more in line with communism than younger cohorts, compared with the same cohort gradient in the West. This holds true for support of the market economy and democracy, preferences for redistribution, and attitudes toward working women. While on average residents in the East express less support for democracy and a stronger desire for redistribution, their preferences for the market economy are on average similar to the ones in the West, and their support of female labor force participation is even lower. The

[11] Again, the online Appendix confirms these findings in a more rigorous way in a regression framework.

latter result could potentially be explained by preexisting differences in preferences between the East and the West.

Thus, we find evidence for long-lasting effects of communism on preferences, despite the fairly rapid adjustment of institutions. This could be one explanation for the comparatively low happiness in the former communist countries despite their economic advancement. Guriev and Zhuravskaya (2009) show that life satisfaction in Eastern European countries is significantly lower than in comparison countries of a similar economic development, with larger differences for older cohorts.[12] The disconnect between preferences and (macro) developments could also contribute to an explanation of political setbacks in countries of Eastern Europe. If individuals in these countries on average prefer large-scale measures of redistribution, yet the actual economic system is not in line with those preferences, people might cast votes for more extreme parties. The fact that support for democracy in Eastern Europe is lagging behind Western European countries may further help explain the support for strong leaders in some Eastern European countries. The long-term effect of communism on preferences could thus still undermine the support for the new institutions.

■ *We thank Mariia Bondar, Carla O. Keller, Paul Reimers, and Penghui Yin for excellent research assistance, and Sergei Guriev and the editors of this journal for very helpful comments.*

[12] While Guriev and Melnikov (2018) conclude that this happiness gap has closed in recent years, partly driven by an average increase in life satisfaction in former communist countries that is due to a change in the demographic composition, Djankov, Nikolova, and Zilinksy (2018) report a persistence of the happiness gap.

References

Alesina, Alberto, and Nicola Fuchs-Schündeln. 2007. "Good Bye Lenin (or Not?) The Effect of Communism on People's Preferences." *American Economic Review* 97 (4): 1507–28.

Bick, Alexander, Bettina Brüggemann, and Nicola Fuchs-Schündeln. 2019. "Hours Worked in Europe and the United States: New Data, New Answers." *Scandinavian Journal of Econ*omics 121 (4): 1381–1416.

Bick, Alexander, and Nicola Fuchs-Schündeln. 2018. "Taxation and Labor Supply of Married Couples across Countries: A Macroeconomic Analysis." *Review of Economic Studies* 85 (3): 1543–76.

Campa, Pamela, and Michel Serafinelli. 2019. "Politico-economic Regimes and Attitudes: Female Workers under State-Socialism." *Review of Economics and Statistics* 101 (2): 233–48.

Cassette, Aurélie, and Sonia Paty. 2008. "Tax Competition among Eastern and Western European Countries: With Whom Do Countries Compete? " *Economic Systems* 32 (4): 307–25.

Djankov, Simeon, Elena Nikolova, and Jan Zilinsky. 2018. "The Happiness Gap in Eastern Europe." *Journal of Comparative Economics* 44 (1): 108–24.

Estrin, Saul. 1991. "Yugoslavia: The Case of Self-Managing Market Socialism." *Journal of Economic Perspectives* 5 (4): 187–94.

Estrin, Saul. 2002. "Competition and Corporate Governance in Transition." *Journal of Economic Perspectives* 16 (1): 101–24.

European Bank for Reconstruction and Development. 2016. *Transition for All: Equal Opportunities in an Unequal World. Transition Report 2016–2017.* London: European Bank for Reconstruction and Development.

Foa, Roberto Stefan, and Yascha Mounk. 2016. "The Danger of Deconsolidation: The Democratic Disconnect." *Journal of Democracy* 27 (3): 5–17.

Fuchs-Schündeln, Nicola, and Tarek A. Hassan. 2016. "Natural Experiments in Macroeconomics." In *Handbook of Macroeconomics*, Vol. 2a, edited by John B. Taylor and Harald Uhlig, 923–1012. Amsterdam: North-Holland.

Fuchs-Schündeln, Nicola, and Paolo Masella. 2016 "Long-Lasting Effects of Socialist Education." *Review of Economics and Statistics* 98 (3): 428–41.

Fuchs-Schündeln, Nicola, and Matthias Schündeln. 2015. "On the Endogeneity of Political Preferences: Evidence from Individual Experience with Democracy." *Science* 347 (6226): 1145–48.

Fuchs-Schündeln, Nicola, and Matthias Schündeln. 2005. "Precautionary Savings and Self-Selection: Evidence from the German Reunification 'Experiment.'" *Quarterly Journal of Economics* 120 (3): 1085–1120.

Guriev, Sergei, and Nikita Melnikov. 2018. "Happiness Convergence in Transition Countries." *Journal of Comparative Economics* 46 (3): 683–707.

Guriev, Sergei, and Ekaterina Zhuravskaya. 2009. "(Un)Happiness in Transition." *Journal of Economic Perspectives* 23 (2): 143–68.

Kornai, János. 1990. "The Affinity between Ownership Forms and Coordination Mechanisms: The Common Experience of Reform in Socialist Countries." *Journal of Economic Perspectives* 4 (3): 131–47.

Lippmann, Quentin, Alexandre Georgieff, and Claudia Senik. Forthcoming. "Undoing Gender with Institutions. Lessons from the German Division and Reunification." *Economic Journal*.

Malmendier, Ulrike, and Stefan Nagel. 2011. "Depression Babies: Do Macroeconomic Experiences Affect Risk-Taking?" *Quarterly Journal of Economics* 126 (1): 373–416.

Pop-Eleches, Grigore, and Joshua A. Tucker. 2017. *Communism's Shadow: Historical Legacies and Contemporary Political Attitudes.* Princeton, NJ: Princeton University Press.

Svejnar, Jan. 2002. "Transition Economies: Performance and Challenges." *Journal of Economic Perspectives* 16 (1): 3–28.

Journal of Economic Perspectives—Volume 34, Number 2—Spring 2020—Pages 192–214

The Basic Economics of Internet Infrastructure

Shane Greenstein

This internet barely existed in a commercial sense 25 years ago. In the mid-1990s, when the data packets travelled to users over dial-up, the main internet traffic consisted of email, file transfer, and a few web applications. For such content, users typically could tolerate delays. Of course, the internet today is a vast and interconnected system of software applications and computing devices, which society uses to exchange information and services to support business, shopping, and leisure. Not only does data traffic for streaming, video, and gaming applications comprise the majority of traffic for internet service providers and reach users primarily through broadband lines, but typically those users would not tolerate delays in these applications (for usage statistics, see Nevo, Turner, and Williams 2016; McManus et al. 2018; Huston 2017). In recent years, the rise of smartphones and Wi-Fi access has supported growth of an enormous range of new businesses in the "sharing economy" (like, Uber, Lyft, and Airbnb), in mobile information services (like, social media, ticketing, and messaging), and in many other applications. More than 80 percent of US households own at least one smartphone, rising from virtually zero in 2007 (available at the Pew Research Center 2019 Mobile Fact Sheet). More than 86 percent of homes with access to broadband internet employ some form of Wi-Fi for accessing applications (Internet and Television Association 2018).

It seems likely that standard procedures for GDP accounting underestimate the output of the internet, including the output affiliated with "free" goods and the restructuring of economic activity wrought by changes in the composition of firms who use advertising (for discussion, see Nakamura, Samuels, and Soloveichik

■ *Shane Greenstein is the Martin Marshall Professor of Business Administration, Harvard Business School, Boston, Massachusetts. His email address is sgreenstein@hbs.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.2.192.

2016, or the Spring 2017 symposium in this journal with articles by Feldstein 2017; Syverson 2017; Groshen et al. 2017). To illustrate the magnitude of the measured economic changes, online advertising contributed $105.9 billion in revenue to the GDP in the categories of Internet Publishing and Broadcasting as well as Web Search Portals in 2017, which had grown 250 percent in the previous five years. The Census Bureau estimates electronic retailing at over $545 billion for just electronic shopping and mail order houses (NAICS 4541), a growth of 65 percent over the same period (based on the Census data on Statistics of US Business).

The external face of the internet has become part of everyday life. However, the internal structure and operation of the internet have remained largely invisible, both to the public and to most economists. This essay will begin by discussing the processes that support delivery of internet services. The internet's infrastructure contains many different types of equipment: root servers, fiber, broadband lines, networking switches and routers, content delivery networks, cellular towers, and others. Meanwhile, the internet's "backbone" consists of enormous data lines, specifically, the lines that interconnect networks and core routers for transmitting packets of data. Other elements of the internet infrastructure include cloud facilities and the parts of the internet that have been taken inside large firms like Google and Amazon. With an understanding of the mechanics of the internet structure, it becomes possible to address questions like: What determines the pricing and terms for exchanging data? What determines the incentives for improving infrastructure? How evenly spread is frontier digital infrastructure across regions?

The discussion will illustrate some classic issues in the economics of networks. Networks which have an agreed upon set of standards and rules can be self-perpetuating in a wide range of circumstances because existing users and potential new ones will be attracted to the well-established network. However, when a network involves both multiple end-users and multiple players within the network who can impose costs and fees on each other, there may be times when negotiations may threaten to deadlock. Expanding a network in its original form may be fairly straightforward, but more complex changes to the operation of a network can be problematic, both because such changes may threaten to disturb the shared rules that make the network function and because the players who would need to invest in the change may find that they are not able to recoup a sufficient share of the benefits from other players in the market to make it worthwhile.

The discussion will focus on practices in a North American context and will oversimplify the explanations of its engineering. However, it should generate an understanding of how internet infrastructure works as well as it does. Technical terms will be introduced and explained as they arise. Additionally, Table 1 provides a glossary.

How Does Internet Data Travel?

To understand how the internet connects so many devices, let's start with a basic example: how a single user request for information from, say, Wikipedia, generates

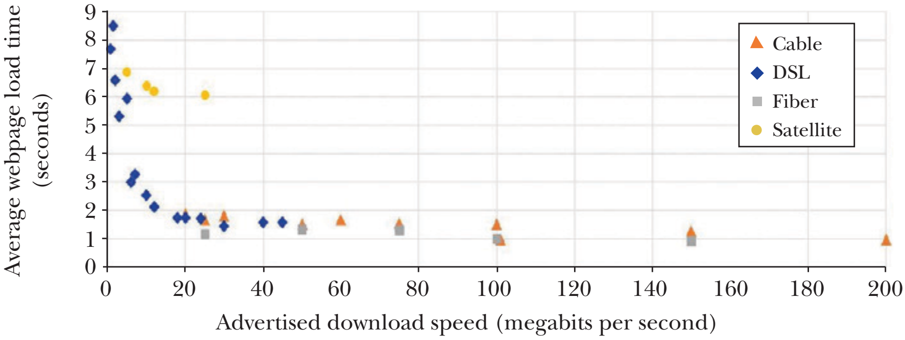
Table 1
Glossary of Some Internet Terminology

Term	*Definition*
Backbone	The long distance and high capacity routes between interconnected networks and core routers in the internet.
BGP	Border Gateway Protocol. The most commonly used protocol for routing traffic on the internet and is one among many governing how network switches and servers send packets of data through the network. The most recent draft dates to 2006. See https://tools.ietf.org/html/rfc4271.
Broadband	Any high-speed internet access that is always on and faster than dial-up access.
CDN	Content Distribution Network. A distributed system of computers that acts as an intermediary for original content, and delivers content transparently to end users.
Cloud computing	An evolving model for enabling a ubiquitous and on-demand shared pool of configurable computing resources. Users typically provision these quickly.
Collocation facility	A location in which servers and other computing hardware reside.
DNS	Domain Name System is set of naming and numbering rules for affiliating common words with IP addresses, consistent with TCP/IP. Today ICANN oversees the system used on the internet.
DOCSIS	Data Over Cable Service Interface Specification. Developed by Cable Labs for cable system delivery of internet access (see discussion in Knieps and Bauer 2016, and Clark 2018).
DSL	Digital Subscriber Line. A form of broadband access retrofitted to telephone lines.
ICANN	The Internet Corporation for Assigned Names and Numbers. The nonprofit organization responsible for coordinating the maintenance and procedures of several databases related to the namespaces and numerical spaces of the internet. https://www.icann.org/.
IEEE	The Institute of Electrical and Electronic Engineers is a global association and organization of professionals working toward the development, implementation, and maintenance of technology-centered standardized products and services. https://www.ieee.org/.
IP address	Internet Protocol. Every device on the internet must have this numerical label assigned to it.
IXP	Internet Exchange Point. Typically a building operated by carriers or by a third party and configured for carrier colocation and interconnection of data traffic.
Protocol stack	The software that implements a family of protocols. These define a set of rules and regulations that determine how data transmits in telecommunications and computer networking.
TCP/IP	Transmission Control Protocol/Internet Protocol. This packet-switching protocol defines how to assemble packets, defines addresses when networks connect to each other, and is a family of protocols that determines the format and error correction processes for packets of data in the internet.
Wi-Fi	It is *not* wireless-fidelity, but is a set of protocols used by wireless routers and based around an IEEE 802.11 family of standards.

a number of instantaneous actions. This involves an explanation of the mechanics of moving data, typically unseen by the user.

Figure 1

Average Weighted Load Time Compared with Advertised Download Speed Federal Communications Commission (December 2018)



Source: Federal Communications Commission (2018).

Here is a simplified explanation of the mechanics: The user employs a web browser that has been installed on a computer, smartphone, or other web-enabled device. The user has access to an Internet Service Provider, or ISP. ISPs provide wireline or wireless access by building and operating the physical equipment that carries data from one place to another. The internet service provider takes the user's request to a name server. The name-server associates an internet protocol (IP) address with the requested destination—in this example, Wikipedia.org. Thus informed, the user's browser directs the query to the server with that IP address. Wikipedia's server responds by releasing the requested data in packets, which are formatted to comply with a specific protocol used to interconnect devices on the internet. That data travels to the user's ISP, which delivers it to the user's device, where it is rendered by the device into in a form the user can view.

Several different market transactions support this two-way flow of information. First, market transactions visibly determine the behavior of internet service providers who are typically paid on a monthly basis. There are broadly two types of ISPs: wireline and wireless providers. Wireline providers vary in their technology—listed here in order from slowest to fastest: Satellite, Digital Subscriber Line (DSL), cable modem, and fiber. Satellite in geostationary orbits deliver and receive data to and from almost any earthly location fitted with a "dish," which communicates with the satellite. DSL service is a retrofit on top of telephone lines to suit it to carrying data. Cable modem service involves the addition of switches and modems consistent with Data Over Cable Service Interface Specification (DOCSIS), which adds data services to cable television systems. Fiber typically involves newly laid lines of fiber optic wire to the customer.

Figure 1 shows a standardized test conducted by the Federal Communications Commission of several advertised tiers of speeds from 17 companies whose service is representative of the experience of the vast majority of US users. The data rates are expressed in megabits per second and translate speeds into a standardized user experience downloading web pages. As illustrated by Figure 1, users experience different download speeds from different tiers of advertised speeds for access technologies. Monthly prices vary accordingly. Typical satellite services cost $90 to $120 a month, on top of set up costs of at least $300 to $500. For DSL, monthly prices tend to range from $30 to $50 a month for only internet service. The largest provider of DSL services in the United States is AT&T, with approximately 16 million subscribers. Prices for cable modem service range from $50 to $80, depending on speed and data caps. The largest providers of cable modem is Comcast, with over 28 million customers. Prices for fiber to the home tend to range between $40 and $80 per month for only internet, depending on speed and data caps. The largest provider of fiber to premises and homes is Verizon Fios, with approximately 7 million subscribers. In any given location the set of options may be more limited to zero, one, or two wireline providers, plus a potential over-builder.

Wireless options differ in use from wireline broadband. While satellite service is available anywhere, most of its users are in low-density locations lacking wireline providers due to its low expense and speed. Estimates put the number of users at more than 8 million households in the United States. The largest providers of wireless services are the carriers Verizon Wireless and AT&T Wireless, with more than 150 and 160 million subscribers, respectively.

Another set of market transaction is invisible to users. The name-server firms are paid by the owners of domain names or their surrogate parties acting in the interests of website owners. The largest US name servers are Cloudflare, Amazon Web Services, and Akamai. While a name server may be a stand-alone firm, it has become increasingly common to offer name service bundled with other services, such as security. In addition, some organizations that send out large volumes of messages will provide their own in-house name server rather than paying for third-party services (for an explanation of this choice, see Bates et al. 2018).

Five Options for Data to Travel

With internet service providers and name servers playing their roles, one crucial step remains: how does the data actually travel between the internet service providers of the user and a content provider like Wikipedia and vice versa? Internet data can follow a multiplicity of paths between two points, which gives the system immense flexibility. How is the route for any given message determined? All options make use of the same routing tables and software protocols, which typically direct the packets of data to the least congested route. That process involves what is largely an engineering decision about how all networked participants collectively must behave in the presence of congestion on some routes. However, we will defer an explanation of how the prices for sending data are determined until later in the paper—because understanding the economics is more easily done after an explanation of the network's mechanics. For now, we will focus on the path taken by data as it goes from user to Wikipedia and back again. The data can travel between user and content provider by one of five options.

The first option is the simplest. If the user and the server contract with the same internet service provider, such as Comcast, then the data can be requested and delivered within the network of a single ISP. This path is common for bilateral communications between individuals, such as electronic mail—the majority of which involves two closely located participants. However, most other traffic, particularly traffic to support web and streaming applications, tends to involve content providers and recipients in far apart locations. Those interactions do not tend to stay within a single network due to the geographically fragmented and unconcentrated provision of US internet services providers.

That brings us to the most common current option in which the internet minimizes delays by rerouting a user request from server to content delivery networks (CDNs), which are geographically distributed networks of servers located near end users. Sometimes this is called "moving data to the edge of the network." Because such networks are physically close to users, CDNs reduce the response time. Many content providers choose to cache content at the CDN and update only the most timely and popular content so that most users are, in effect, exchanging content with the CDN rather than the ultimate provider of content. CDNs also can provide a layer of reliability and security: for example, even when some servers have gone down, the cached content in a CDN may keep a firm's content available for users. CDNs can also buffer content from a "denial-of-service" attack (in which an attacker seeks to disable a target by flooding it with messages).[1]

Content delivery networks did not exist at the outset of the commercial internet, but today, almost every commercial participant of any size employs them in some way for popular content. The largest third-party provider of CDN services in the United States is Akamai, with revenues of $2.7 billion in 2018. The next-largest provider of such services, Cloudflare and Limelight, had revenues of $192 million and $184 million in 2018, respectively. Though content delivery networks are unseen to the user, the vast majority of data received by a user comes directly via this route.

Three other options for moving data have been used for over two decades since the privatization of the internet (Greenstein 2015), but it is difficult to derive estimates of their frequency of use. In the distant past, all were more commonly used to move data from content providers directly to users—that is, without the intervention of a content delivery network. Today, these same forms move data from content providers to CDNs, complementing the CDNs in the vast majority of requests. These three forms—private peering, internet exchange points, and transit carriers—act as a substitute for the CDN in a smaller set of cases, as when the user requests unpopular content, or the content provider does not make an arrangement to employ a CDN.

[1] Readers may be interested in Patent 8613089B1, *Identifying a Denial of Service Attack in a cloud-based proxy service,* assigned to Cloudfare at https://patentimages.storage.googleapis.com/a0/90/f7/3f8aa8ef076cf4/US8613089.pdf.

"Private peering" arises when Wikipedia and the user (and the CDN supporting the user), have different internet service providers, but those two providers have a direct point of contact, and made a bilateral contract with each other to govern the exchange of data. In a typical contract, no money changes hands if over a month their data flows back and forth in rough proportion to each other. If one party gives a higher proportion of data to the other, the carrier who gives more data (on net) pays the other carrier for taking the traffic. Typically, these payments arise when traffic exceeds a negotiated ratio between four-to-one or eight-to-one—but no simple sentence can describe these contracts and negotiations (for discussion, see Norton 2014).

Two or more internet service providers also may exchange data at an internet exchange point (IXP), which may be run by a separate organization and configured as a place for carriers to meet and interconnect so they can exchange traffic. Each carrier pays a fee to the organization that houses the equipment that facilitates exchanging the data and may make numerous investments in the structures, backup energy, and equipment to keep these operating under all circumstances. Unlike private peering, all participants generally agree to send and take whatever volume of data their connection's capacity can handle. Charges may vary for each tenant and often has no relationship with volume of traffic. There are hundreds of IXPs in the United States and more all over the globe. The largest operator is Equinix, with over $5 billion in revenue and over 200 data centers in many cities, with some of these configured to serve as IXPs.

When the internet service providers for the user and for Wikipedia in our base example do not have any direct contact with each other, not even via an internet exchange portal, then a last possible form of making contact arises. One or more networks' lines acts as a transit carrier between the two ISPs. The carriers providing transit may have received compensation for that action depending on all their contracts with other carriers.

Incentives for Investment, Expansion, and Improvement

Notice an economic implication of this system: carriers have incentives to build more lines, make more connections, and relieve congestion, if and when it helps the firm to gain revenue or to avoid charges from other firms. Internet service providers face additional incentives to increase capacity and make connections if it enables them to increase revenue from users and/or avoid operational costs. These incentives appear to be consistent with a desirable long-term outcome— namely, more efficient and better options for routes to send and receive data. An interesting open question concerns the size of the private incentives in relation to the gains to the network. Transit lines are one component in a system, and improvements in one component confers benefits to all the other complementary components. Do most of the gains from better transit lines go to the content providers who use them, to the users who enjoy previously slower content, or to the internet service providers who may gain revenue from users for better services? The answer partly depends on pricing, which we discuss later.

A related question arises about the incentives to install content delivery networks. A third-party commercial content delivery network negotiates interconnection with an internet service provider or wireless access provider for the right to "collocate" a server close to users. The ISP or another network provider also may charge a "transit" fee to the CDN to take data over its network lines (that is, from the content firm's servers to the equipment installed by the CDN). The original content providers pay the CDN provider to redistribute content to users from the CDN's servers, which the content provider updates at an arranged schedule over the course of the day. This contractual arrangement arises in virtually any, albeit the smallest, ISPs in the United States, which suggests it serves the interest of ISPs.

Some large content providers, such as Google, Apple, Microsoft, Facebook, Amazon, and Netflix, operate their own content delivery networks and tailor the technical features to their own applications and services. Again, they negotiate a price that they pay to internet service providers for "collocation," and they sometimes pay fees for data transit. In practice, only large firms opt for this action because it is usually less expensive to contract with a third-party CDN for small to medium volumes of traffic. Also, for a number of reasons—scaling issues, negotiating frictions, and the collocation expense—some firms prefer to locate some of their private CDNs at internet exchange points, not within internet service providers.

It is an open question: Do most of the gains from better content delivery networks go to CDN providers who operate the servers, to the content providers who use them, to the users who enjoy previously unobtainable content, or to the internet service providers who charge collocation fees and also may gain revenue from users for better services? As with any network component, it is unclear how the private incentives compare with network-wide gains.

This question is important because the rise of content delivery networks was both a cause and symptom of changing user needs and dramatic network improvements. Many users have migrated to broadband with higher bandwidth, which increases user speeds. These users are more likely to desire and support new applications, which would have been infeasible without CDNs, such as "over-the-top" streaming services such as Netflix, Sling, Disney+, or HBO Go—that is, services that bypass cable or satellite television content and instead are provided directly to consumers over the internet.

The dramatic improvements are most visible in the heavy evolution of applications of the internet and the traffic that accompanies them. In the earliest days of the internet, text dominated traffic either in the form of email or passive browsing. By modern standards, the volumes of data were small in either direction. In contrast, households today receive increasingly many more magnitudes of data than they send, as the majority of traffic that households receive changes from static content to video and streaming (Huston 2017). For example, back in 2013, a median household used 20–60 gigabytes of data per month (Federal Communications Commission 2013). However, streaming a standard or high definition movie generates between 1 and 3 gigabytes per hour, far more data than any passive web-browsing ever could generate.

Merely binge-watching a single streamed series could massively increase household data use. Meanwhile, the largest streaming service, Netflix, has increased its US subscribership from 20 to 60 million over the second decade of the century, and it is far from the only streaming service. In short, as streaming of television and movies rises in households, the capacity of the underlying infrastructure to handle data-intensive applications must increase. It is always hard to answer the question of whether incentives to invest are optimal, but the experience of the internet certainly suggests that private incentives to invest have been sufficient to produce a dramatic expansion and upgrading of network components.

Data Centers and the Cloud

At the outset of the commercial internet, virtually all firms housed their servers on company premises. Businesses, however, eventually learned to gain scale economies by consolidating computing resources in one location, which gave birth to the data center. These structures contain many rows of servers on racks, matched to routine operations for support and maintenance of the internet. Designers eventually learned to configure these structures to house massive numbers of servers devoted to storage or computation, using architectural features that encourage low energy use and ensure reliable operations in the event of emergencies, among many features.

Some of the different ways that the market can send and receive information, such as peering and interconnection, also occur at some data centers configured for such a purpose. The inside wiring of a data center may support a specific set of activities. The data center for the New York Stock Exchange, for example, is located in New Jersey, and it permits many firms to access trading services at especially fast rates. As another example, a segment of business users in health, finance, and transportation require high security and high reliability—that is, 99.99 percent uptime—especially in critical functions that support transactions with sensitive customer data. These data centers may contain expensive backup generators, expensive structures to prevent flooding, and reinforcements in the floors to reduce any vibrations from passing vehicles. These expensive features pay off in certain situations; for example, due to built-in resiliency and smart site-selection, the data centers in Houston continued operating without interruption during and after the flooding of Hurricane Harvey in September 2017.

Small data centers house tens of thousands of servers and can cost more than $100 million to build from scratch, while large data centers house hundreds of thousands of servers and can cost several billion dollars to build from scratch. One of the largest third-party facilities in the United States, the Lakeside Technology Center, resides two miles south of downtown Chicago in 1.1 million square feet of a converted building that formerly housed R.R. Donnelly's printing facilities. It is owned by Digital Realty Trust, a holding company that manages more than 200 data centers around the globe, generating just over $3 billion in revenue in 2018.

This building is an exception to the norm for data centers, which are typically large one-story buildings built on an expanse of land near abundant, inexpensive electricity and high-quality interconnection with the internet, often at a suburban location not far from the business users. The largest agglomeration of data centers in North America is in Ashburn, Virginia, just outside Washington DC, near Metropolitan Area Exchange, East (commonly referred to as MAE-EAST), which is one of the oldest IXPs in the United States.

Contracts for data centers cover every conceivable arrangement between ownership and rental markets. At one extreme, many buyers with generic needs—such as storage for backup—rent data center space, own the servers, and let others manage the building. At the other extreme, firms with unique computing needs—such as Facebook, Apple, Microsoft, Amazon, and Google—own and operate large private data centers and configure the building and servers to suit their applications.

A "cloud" service involves a data center that rents its services for storage, computing, or their respective applications to a service such as database, with the additional feature that users can turn the service turn off and on at will. The major cloud providers also increasingly offer additional software services for a nominal charge or none at all. For example, Amazon Web Services offers scores of cloud software services. Microsoft Azure supports many Microsoft products, such as Outlook, as a cloud service. Google offers Tensor Flow, a standard tool for machine learning, at no charge with its cloud service.

The demand for cloud services has grown as the services have improved and declined in price. Byrne, Carrado, and Sichel (2018) estimate a quality-adjusted price decline between 2009 and 2016 at 17.3 percent per annum for Amazon Web Services. Estimates of the growth in expenditure and market share within the industry depend on the precise definition of sales (for discussion, see Byrne, Carrado, and Sichel 2018; Coyle and Nguyen 2018), but some of the three biggest players are those just mentioned: Amazon Web Services (AWS), Microsoft Azure, and Google Cloud. In 2019, for example, AWS was widely regarded as the largest of these three cloud providers and brought in $35 billion of revenue—an increase of 40 percent from the prior year. The others are also growing rapidly. The appeal of cloud facilities comes from their flexibility, wide range of tools, and the option to substitute variable costs for fixed ones (Wang and McElheran 2017). It has enabled experimentation by many entrepreneurial applications (Ewens, Nanda, and Rhodes-Kropf 2019).

The private cloud providers increasingly use complex architectures to balance the loads from user demands—for example, using a mix of data centers for high-scale tasks and content delivery networks for rapid response for timely content. Cloud facilities provide updates to the CDNs at timely intervals and secondary response of less popular content, while home servers provide updates at slower intervals and respond to requests of the least popular content. These may shift their loads as peak user demand shifts over the course of the day across different geographic areas.

When Large Firms Operate Their Own Internet Infrastructure

Many large firms in application markets and platforms, such as Microsoft, Apple, Alphabet, Amazon, and Facebook, operate their own internet infrastructure rather than use third-party market suppliers. For example, all of them operate their own data centers and content delivery networks. As another example, Alphabet/Google connects its own data centers to each other with its own backbone lines and thus bypasses backbone lines it could lease from network operators. Large firms that integrate into complementary functions presumably do so because they can operate processes at a lower cost than third-party providers offer. It also may help achieve higher performance once the processes are tailored to specific needs. In the case of Google, for example, the lines help balance loads across its many data centers and CDNs over the course of the day. Large firms also may find scope economies across multiple related services, thereby spreading the efficiencies, or enabling them to offer services as bundled offerings that appeal to users (Bates et al. 2018). For example, as part of a suite of security offerings to protect content, Cloudflare offers CDN and name-server services inside one package of many services.

When large firms bring internet infrastructure in-house, the effects for the network as a whole can be positive, neutral, or even negative. For example, several of the largest firms that operate large data centers—like Microsoft, Google, and Amazon—began offering cloud services some years ago. Users became accustomed to the resulting efficiencies, and demand for these services is growing rapidly. The network economy, thus, benefited from the entry of these firms into the supply of cloud services.

However, the experience of Google Fiber illustrates another type of situation. Google started a new division to offer high-speed fiber to households and entered several cities with contracts for television, telephone, and internet service. While commercially successful in several cities, as of this writing, this division has paused its investments while seeking to overcome some challenges.[2] So far, therefore, the visible gains have been modest and localized to the few places where entry has been built or, at best, demonstrative of what might be possible elsewhere. Even if Google Fiber does cover all its intended cities, it will cover no more than 10 percent of the US population.

When large firms integrate into internet infrastructure, some outcomes of potentially greater concern arise; providers of complementary services must negotiate with large dominant firms, and thus potentially face contract terms they would not have encountered in a competitive setting with more options (Rogerson 2018). Also, there is a long-standing concern that increasing use of proprietary processes inside the largest firms can diminish the likelihood of generative innovations that

[2] As of this writing, Google Fiber offers service in Salt Lake City, NV; Provo, UT; Kansas City, MO; Austin, TX; Nashville, TN; Charlotte, NC; Atlanta, GA; Raleigh-Durham, NC; Orange County, CA; Huntsville, AL; and San Antonio, TX. It entered and exited Louisville, KY. Google Fiber has plans and permits to enter at least a dozen more cities, but no announced timeline.

would have arisen with wider use of open protocols (Zittrain 2008). These concerns play a significant role in antitrust or regulatory analysis. Therefore, an important open question for debate concerns the degree of market power and range of circumstances over which these concerns apply.

The rise of private data centers and the cloud, once again, raises important economic questions about competitive behavior and private incentives from improvements in networking infrastructure. What are the distribution of gains between users and producers from improvements in one part of a network, such as the cloud? In the presence of the gains shared by users, are competitive incentives sufficient? Do they favor some users over others? What are the long-term competitive prospects for new entry by entrepreneurial firms who use third-party services? These are important open research questions.

Protocols and Governance

Protocols are a set of rules and regulations that determine how data makes it through the network. A networking protocol defines conventions for processes, which include definitions for both the format of data packets, as well as for recovery in the event of transmission errors. For example, the TCP/IP (Transmission Control Protocol/Internet Protocol) is a family of protocols that sets a format for packets of data in the internet, defines addresses when networks connect to each other, defines how to assemble packets of data that arrive through the internet by different routes, and includes error correction processes. The BGP (Border Gateway Protocol) is the most commonly used protocol for routing traffic on the internet, although it is just one among many that governs how network switches and servers send packets of data through the network.

Engineers say equipment is compatible with other equipment only if both sets have adopted the same protocol. Each protocol lives with many complementary protocols in a protocol "stack"—a family of related protocols assembled together. The protocol stack acts as a reference model for designers, who largely aspire to make compatible equipment (for additional descriptions, useful starting points are Clark 2018; Knieps and Bauer 2016; Greenstein 2015).

The protocol stack for the internet (mostly) sends data packets along the least-congested route, a feature that delivers data quickly even when there are many potential routes for data and bottlenecked capacity along points of the network. This feature has become increasingly important because many modern internet applications, such as gaming and streaming, depend on fast delivery of data.

Infrastructure firms and carriers largely comply with protocol stacks; after all, they can offer profitable services while doing so. This outcome should not be taken for granted. It represents a notable departure from a prior era of practices, where many different firms offered proprietary protocols and networks and these did not interoperate. Since the birth of the commercial internet in the middle of the 1990s, however, compatibility with the internet protocol stack has been self-reinforcing. Compliance with protocols by all other suppliers and users further motivated widespread adoption and persistent use of these protocols by any participant, and it motivated development of many additional innovative services built on top of this equipment. The incentives for continuity are apparently strong in the modern internet, in spite of variance in the cause and size of the network effects across participants.

It is possible for situations to arise in which a break with the existing protocols makes sense to a decision-maker. Remarkably, none of those pressures has been sufficient in recent decades to generate stark breaks with internet protocols, though there are examples of partial movement in that direction (Simcoe and Watson 2019 provide a useful framework). For example, operators of the "dark web" prefer not make their content searchable, because they (allegedly) support illegal activities, such as the exchange of pirated material. Network effects also may not operate at the international level as different governments adopt mutually incompatible practices for their domestic networks, in some cases to censor content, but also to impose limits on the operations of applications consistent with local preferences for privacy, security, copyright, and other government policy. Some of these actions have begun to migrate into the infrastructure layers, where governments impose, for example, packet-inspection processes in routers, or back-door design within operating systems to permit surveillance. These actions and policies frame open questions about the risks of losing seamless interoperability, or "splintering" the internet. These topics deserve attention from economic researchers.

These observations also motivate questions about the governance for improving protocols. For the most part, nonprofit organizations design and upgrade the protocol stack used for internet infrastructure. For example, the Internet Society provides the home for the Internet Engineering Task Force (IETF), which governs the protocols behind TCP/IP and BGP, and many others. The Internet Corporation for Assigned Names and Numbers (ICANN) governs assignment of domain names and updates the routing tables used by every switch and router on the internet. A routing table contains information about the topology of a network, and provides guidance about where data packets should go. In modern systems, the tables learn about congestion and send data on routes to avoid the congestion (for discussion, Clark 2018). The Institute of Electronic and Electrical Engineers (IEEE) convenes committee 802.11, which supports the standard underlying Wi-Fi, as well as other technical standards. These organizations convene groups that design, maintain, and upgrade the protocols, and they subsequently charge little for their use. Most also put few legal restraints on how the private sector operates the equipment that uses those protocols.

Many voices influence and determine the actions of the organizations who govern protocols. Given the private stakes, it is no surprise that debates about policies for intellectual property receive considerable attention today, as do debates for criteria about what administrative process should be used to create a protocol. For an example of such a debate, consider the problem of exhaustion of available

IP addresses, which necessitated a redesign of IP addresses to enable growth into the future. Version 6, abbreviated as IPv6, emerged from a debate at the Internet Engineering Task Force. It has been slow to diffuse since it became available. Many blame the new design, which is cumbersome to adopt.

As another example, consider the vociferous debate surrounding the expansion of top-level domain names by ICANN. The internet was designed with 248 country codes, but six domains inside the United States, where no country code was required—com, org, net, edu, gov, and mil—became widely used, especially com. In response to complaints about the limitations arising from the concentration of names under com, ICANN expanded the number of domains to over 1,000, including icu, top, xyz, site, vip, and online. For histories of these and related organizations, and an analysis of their origins, interested readers might begin with Mueller (2004), Simcoe (2012), Russell (2014), Greenstein (2015), and Clark (2018).

The choice of protocols and changes to protocols resembles a public good problem because virtually all users have the same experience, and they can neither opt out nor be excluded from changes. Considerable effort goes into the design of protocols, but not all of them receive equal use. Development of economic theory for when it is worthwhile to change protocols gradually or dramatically, or whether to abandon them at all, is essential for understanding the continuance of the commercial internet.

Pricing and Incentives

For most internet users, the up-front price they face involves a fee from their internet service provider. The vast majority of business users in urban and suburban areas contract for broadband internet access (for the diffusion of broadband, see Ryan and Lewis 2017, Pew Research Center 2019). From 2012 to 2017, payments for access to wireline forms reached $88.7 billion, growing more than 30 percent. Wireline access also became faster, as much as doubling in speed between 2011 and 2018. Payments for access fees to wireless service also reached over $90 billion, an increase of 57 percent (according to Census data from the Statistics of US Business). The revenue increase during this period did not largely arise from a rise in the number of households because most US households already had internet service in 2012.[3]

The market for supply of broadband services has a moderate degree of competition. This supply structure emerged after the replacement of dial-up with broadband as the primary method of internet access (Greenstein and McDevitt 2011). In 1995, virtually all internet access occurred over dial-up; whereas today, approximately 80 percent of US households have broadband internet access in

[3] For example, the Netflix ISP Speed Index comparisons of measured speeds over 2012-2018 yields a doubling of realized speeds for most networks (https://ispspeedindex.netflix.com/country/us/). From 2011 to 2018, only 3 to 5 percent of US households first began using broadband internet, depending on the survey (Pew Research Center 2019).

their homes. Downtown locations in high-density settings experienced greater entry, aimed at business customers and/or multi-occupation residences (Chen and Savage 2011; Connolly and Preiger 2013). Most households in urban and suburban settings have access to at least one or two providers of wireline access, and multiple wireless providers (Wallsten and Mallahan 2013). The typical providers for households are the local telephone company, who typically offers DSL service, and the local cable television provider, who offers data services using modems compatible with DOCSIS, the Data Over Cable Service Interface Specification. In some areas, a third-party "over-builder" may offer fiber to the home, and one local telephone company, Verizon, offers fiber to homes in some of the territories in which it operates. Business in dense urban locations may have access to even more providers.

Some of the many components that play a role in limiting entry of internet access providers include financial reasons, such as high capital costs; regulatory factors, such as limited rights of way, rules raising the costs of over-builders, and laws preventing entry from municipal providers (for example, Seamans 2012); and behavioral dynamics, such as the unwillingness of incumbent firms to enter each other's established territories. In addition, technical forces make some forms of broadband access, such as DSL or 5G wireless service, less effective outside of dense locations (for example, Destafano, Kneller, and Timmis 2018), or make cable service cost-prohibitive. Satellite services remain viable in most terrain, providing a baseline level of service for less-dense areas (Boik 2017). Relatedly, a robust market for supplying cellular towers to support carriers' antennae enables service from two to four providers in all but the least dense locations.

Meanwhile, measured price levels for access have changed little since broadband became the dominant delivery mode for households. The Consumer Price Index (available at US Bureau of Labor Statistics 2020) provides a measure of broadband prices in the price series for "Internet services and electronic information providers" (US city average, all urban consumers), which rises from 73.4 to 77.1 from 2007 to 2019, an increase of 5 percent. The closest comparable index for wireless services (which covers data and also includes the price of telephone calls) shows the Consumer Price Index for "Wireless telephone services" (again, US city average, all urban consumers) dropped from 64.5 to 46.4 over the same period, a decline of 22 percent. In light of the enormous changes in those years—for example, the rise of Web2.0 businesses, the growth of social media, and the explosion of short and long form video and streaming—it is likely that measured prices miss important aspects of the typical user experience.

What is missing? For one, neither index adjusts the price of internet access for the quality of that service. In addition, neither accounts for changes in the quality of ad-supported "free" content (for an approach to the latter, see Byrne and Corrado 2019). Lastly, large growth in access revenues with small growth in the number of subscribers indicates many households increased their expenditure on internet access by moving to higher tiers of service at higher prices. Standard methods for price measurement do not count such migration across tiers necessarily as a change in prices.

When do wireline and wireless services substitute for each other, and when do they complement each other? No general answer exists. Any answer changes over time as access capabilities improve and as modal applications change, and it varies by location of the user. In some applications today, wireline and wireless delivery, such as electronic mail and passive browsing, can substitute when users can tolerate delays. These modes of delivery do not substitute in other applications, such as data-intensive streaming and gaming, where delays interfere with user experience. Sometimes they complement each other, such as when entertainment firms encourage tweeting during an online gaming event or streaming of content. These are difficult questions for a substantial number of households that get their internet through only a wireless smart phone or satellite. As access to frontier infrastructure improves, the extent of substitution and complementarity between wireless and wireline services is an important open topic of research.

Contracts between users and access firms also changed over time. In the earliest years, most access involved a monthly charge and no limitations on use. Greenstein (2015) discusses the disappearance of price discrimination based on the amount of time online. Today price discrimination based on usage of data, combined with data caps, is common in both wireline and wireless contracts. Moreover, wireline and wireless data contracts do not take the same form. Burnham et al. (2013) provides early census of the use of tiered pricing and caps based on the usage of data in wireline. Recent studies show that some users are sensitive to the charges affiliated with reaching a data cap, but they also endogenously select into capacity consistent with their use; for example, those who practice heavy streaming choose plans that allow this without large cost increases (for example, Nevo, Turner, and Williams 2016; McManus et al. 2018).

The incentives for improved internet capabilities and access receive considerable attention from policy analysts. On the one hand, there is the general belief that improvements in the speeds of wireline and wireless access benefit more participants than just the firm providing this access. While access providers potentially gain more revenue, users gain better service, and application providers face the option of a new frontier for their data-intensive services. Once again, the gains from improvement are widespread, while the private costs and commercial risks are concentrated in the one investor—in this example, access providers. As mentioned earlier, an important open question concerns the gaps between private and social incentives to upgrade wireline broadband. Nevo, Turner, and Williams (2016) suggest the gap is substantial, consistent with the presence of insufficient private incentives to upgrade quality at a rate in line with society's broader interest.

Estimating these incentives remains an open research area, especially in upgrades to wireless technology. For example, cellular telephony migrated to new generations of technologies, from 3G to 4G. 4G is the fourth generation of broadband cellular technology, succeeding 3G. 4G uses only packet-switching technology, unlike 3G, which used both packet-switching and (in parallel) the (old) circuit-switching technology. As of this writing, 5G contains much more capacity than 4G, and has only just begun to deploy. In summary, users have increased the use of data substantially on wireless modes as it has deployed, and suppliers have invested in order to support those increased volumes. Were incentives for this evolution too high or too low, and how would answers to this question inform expectations about the ongoing upgrade to 5G?

Another quality-related concern touches on competitive issues. If wireline broadband firms carry video-on-demand, then questions arise over conflict of interest in carrying other forms of internet traffic, which in turn would effect investment, interconnection, and pricing (Rogerson 2018). One other change also may have shaped incentives in the recent experience, and may do so in the future. At the outset of the commercial internet, internet service providers did not charge for accepting data delivered to them to be sent to their direct customers, but today some do. This provides additional revenue for internet service providers, and it has been met with resistance from other providers who interconnect with ISPs because it raises the costs of providing data services over long distances and content delivery networks.

Less data is available concerning the prices and fees that happen behind the scenes in interconnection. Interconnection networks often reach agreement without disclosing terms. Evidence from Zhuo et al. (2019) shows interconnection agreements growing everywhere, with some variance across different geographies due to economic development. Negotiation plays an important role in shaping fees for private peering and content delivery networks. Any time that price negotiations take place in the shadow of alternative options for accomplishing functionally equivalent outcomes, then those options discipline attempts to raise prices or exploit negotiating advantage in other ways. Conversely, infrastructure firms have a negotiating advantage when they provide services for which there are no substitutes, and/or when they can bottleneck the aspirations of other network participants.

In this setting, are prices inside the internet infrastructure more likely to be determined by a range of competitive options that tend to drive down prices paid by ultimate users? Or are these prices more likely to be determined in many situations of limited competition and bottlenecks? There is limited evidence on these questions. An optimistic view focuses on how much all participants know about the conduct of negotiations (Norton 2014). Often negotiations resolve themselves without incident: in fact, there has not been a prominent example of breakdown in negotiations since late 2013, when Netflix and the four largest providers of internet access in the United States could not reach a negotiated settlement, which led to widespread congestion issues affected streaming speeds and reliability at tens of millions of households (Greenstein and Norris 2015). The underlying issue in this case was that the original model for the commercial internet involved no charges for delivery of data to internet service providers, and attempts to impose such charges played a role in the negotiation breakdown between Netflix and four large ISPs.

Perhaps the most salient evidence for being optimistic that these prices are being shaped and determined by a range of (reasonably) competitive options is the long-term record of a symbiotic relationship between advances in infrastructure, growth in access revenue, and advances in revenue for electronic commerce. Internet infrastructure has contributed to reducing several frictions related to conducting commerce; indeed, Goldfarb and Tucker (2019) argue that digital technology has improved economic activity largely through reduction in these frictions. The earlier illustration of a transaction focused on a user's request for data from Wikipedia, a nonprofit organization, but if the user had requested data from profit-seeking firms, for example, it would have triggered additional commercial actions. If there had been advertising, then advertisers would have paid for ad exchanges and geolocation of the IP address, so the user receives a geographically-appropriate advertisement. Related processes may have personalized the advertising further. All of these steps would take place virtually instantly, and largely unseen to the user. Had the user bought or sold a product, many additional processes would have supported fulfillment of the order and would have increased the flow of funds to infrastructure—data centers, switches, and transmission lines—to support the transactions.

A pessimistic view of the internal pricing of the internet points out that the features of any known incident lack transparency, especially in the first decade after the millennium (Greenstein 2010). In the 2013 Netflix incident, for example, households did not know who unreasonably held up whom—the local access provider or Netflix—as performance declined for many applications other than those involved in negotiations. Because the terms for settlement did not become public, competing interpretations of the event remain unresolved, and so too do questions about whether government intervention could have alleviated the decline in internet performance experienced by users.

A pessimist also would observe how the dependence of network participants on many firms creates difficulties in assigning responsibility for inadequacies in the delivery of service. An outage can have widespread consequences. For example, when services like Slack, Quora, and Medium all became unavailable February 28, 2017, users had no way to know that Amazon cloud storage had gone down due to a single maintenance person's faulty actions at one AWS facility in Northern Virginia, which caused a number of Amazon web servers to go offline. As another example, when services such as CNBC, Netflix, and Twitter went down on October 21, 2016, users had no way to know that it resulted from a distributed denial-of-service attack on Dyn, who provided the name-server services for these firms. However, events like these often provide the fodder for debate about legal or regulatory frameworks for denoting who assumes responsibility for compensation from economic loss, and relatedly, whether these frameworks provide sufficient incentives to suppliers for investment in risk mitigation.

Geographic Availability

An uneven geographic supply of internet infrastructure is not the only reason why some areas have high rates of non-adoption of internet access, but it plays a major role. Today approximately 10 percent of the US population does not use the internet (Anderson et al. 2019). Some of that non-adoption is linked to demographic features of users, such as older age, low income, and less education. But an important factor is the location of a household, namely, in a rural or low-density location. While 97 percent of the land in the United States is rural, according to the Census Bureau, 19 percent of the population lives in rural locations—that is, areas with sparse residential housing.

Cutting-edge internet infrastructure tends not to be available in low-density regions. In some of these locations, even internet infrastructure with older technology may not be available (for additional discussion, see Forman et al. 2018). For example, according to a Pew survey conducting in February 2019, 63 percent of US rural residents say they have broadband access, compared with 80 percent for suburban residents (Perrin 2019). Moreover, 20 percent of rural households are wireless-only; for comparison, 25 percent of low-income households are wireless-only (Anderson 2019). "Unavailable broadband service" or "low quality" is cited by 22 percent of non-broadband home users as reasons for relying solely on wireless smartphones for internet service, among those who do. However, the price of wireline broadband access is the most frequently cited reason to go to only wireless, with more than half the users citing high prices as the primary reason.

An uneven geographic supply of internet infrastructure arises for many reasons. The costs of supplying internet service to a given geographic areas may reflect economies of scale: that is, when installing and operating cell towers, data centers, and content delivery networks, these structures endogenously locate near a greater number of densely located users, because it facilitate faster returns on investment. Laying lines between locations involves high fixed costs and low marginal costs, so low density may not have sufficient demand to incentive such investments. A demand for higher quality can also drive unequal dispersion because suppliers prefer to build out initially in more affluent and urban locations where a greater number of buyers are more willing to pay for the expensive frontier quality. Marshallian agglomeration can reinforce these differences, with richer, dense urban locations attracting skilled labor and receiving infrastructure closer to the technological frontier.

Geographic variance in supply potentially creates different experiences across households and businesses in different locations. Econometricians often look for this type of variance. However, many variations arise only at a fine level of geography, such as a neighborhood, and attempts to measure availability at this fine-grained level have encountered numerous challenges. For example, an attempt to create a National Broadband Map, which began in 2011, went through several revisions; it was regarded as accurate in some but not all locations, and was discontinued in December 2018. As of this writing, the Federal Communications Commission is developing a new mapping program.

Public policy in such situations also faces some tradeoffs affiliated with economic factors that defy easy measurement. Many users prefer a local supply of internet infrastructure when it is available. However, it may be cheaper to use remote data centers, cloud storage and/or satellites, and users may be willing to substitute into distant suppliers under a range of circumstances. These consumer preferences are challenging to learn, but they will shape any calculation of how much society is willing to spend to provide basic internet access to areas where it would not otherwise be supported by the market, and how much society is willing to spend to provide a quality of internet service to these areas above the basic level.

Many different programs seek to address these concerns. For example, the 1996 Telecom Act established the e-rate program, which taxed telephone calls to finance subsidies for rural broadband. Today it raises more than $4 billion annually, focusing on developing broadband internet access in costly locations, and making it available to organizations with public missions, such as libraries, schools, and hospitals. As another example, the 2009 stimulus package included $7 billion of subsidies for rural broadband. At a local level, many local governments also try to shape supply. Many insist through cable franchise agreements that cable providers build out into low-income or low-density areas. Programs to address demand also exist, but are less common. As a condition for approval of a merger, for example, Comcast agreed to offer lower prices to qualifying low-income households, and evidence suggests it had an effect on hundreds of thousands of households (Rosston and Wallsten 2019). Given the range of programs, it is no surprise that there are many debates over the effectiveness of subsidies of different sizes and designs.

Final Policy Questions

Internet infrastructure has been improving over the decades in ways that have enabled an extraordinary gain in internet services, which many users were willing to pay for. In turn, this reinforced incentives to do something less visible to most users—namely, to improve digital infrastructure. The virtuous cycle has gone hand-in-hand with growth in access revenues, growth in advertising revenues for free services (such as search and news), and growth in electronic commerce that takes advantage of online shopping.

A big, open question is whether this growth and improvement will continue into the future. Some factors will slow growth, such as saturated adoption of broadband in households and businesses. Other factors may accelerate it, such as the restructuring of online services to take advantage of the cloud and 5G wireless infrastructure. The history of the last few decades suggests that internet architecture contains a remarkable capacity to adapt to changes but, in fact, every change raises novel challenges. Every change alters revenues and costs for different suppliers, and some of these developments generates disputes, such as those that accompanied the early diffusion of streaming. So it is an open question whether certain kinds of future changes will place undue stress on the governance of internet infrastructure and the resilience of its designs.

So far, no consensus has emerged regarding an internet regulatory framework. Instead, policy has evolved alongside commercial internet growth, including formal policy in statute or regulatory orders (like in conditions for mergers), less formal policy found in public statements about general principles (like speeches from the chair of the Federal Communications Commission), along with the willingness of government actors to intervene. For a history of these policies, see Nuechterlein and Weiser (2005), Greenstein (2010), Greenstein, Peitz, and Valleti (2016), Knieps and Bauer (2016), and Cybertelecom.org.

No futurist foresees a lack of opportunity to restructure wireless and cloud services, nor does anyone foresee ubiquitous competitive broadband arising in all locations in the next few years. These and other internet changes are sure to generate open questions and policy debate for the foreseeable future.

■ *The author would like to thank Tim Bresnahan, David Clark, kc Claffy, Chris Forman, Avi Goldfarb, Gordon Hansen, John Laprise, Enrico Moretti, Abhishek Nagaraj, Frank Nagle, Tommy Pan Fang, Bill Rogerson, Alicia Shems, Heidi Williams, and Pai-Ling Yin for extremely helpful conversations and comments. Timothy Taylor deserves special thanks for masterful editing.*

References

Anderson, Monica. 2019. “Mobile Technology and Home Broadband 2019.” *Pew Research Center*, June 13. https://www.pewresearch.org/internet/2019/06/13/mobile-technology-and-home-broadband-2019/.

Anderson, Monica, Andrew Perrin, Jingjing Jiang, and Madhumitha Kumar. 2019. “10% of Americans Don’t Use the Internet. Who are They?” *Pew Research Center*, April 22. https://www.pewresearch.org/fact-tank/2019/04/22/some-americans-dont-use-the-internet-who-are-they/.

Bates, Samantha, John Bowers, Shane Greenstein, Jordi Weinstock, and Jonathan Zittrain. 2018. “In Support of Internet Entropy: Mitigating an Increasingly Dangerous Lack of Redundancy in DNS Resolution by Major Websites and Services.” NBER Working Paper 24317.

Boik, Andre. 2017. “The Economics of Universal Service: An Analysis of Entry Subsidies for High Speed Broadband.” *Information Economics and Policy* 40: 13–20

Burnham, Brad, Shane Greenstein, Neil Hunt, Kevin McElearney, Marc Morial, Dennis Roberson, and Charles Slocum. 2013. *Issues in Data Caps and Usage Based Pricing.* Washington, DC: Federal Communications Commission.

Byrne, David, and Carol Corrado. 2019. “Accounting for Innovation in Consumer Digital Services: IT Still Matters.” NBER Working Paper 26010.

Byrne, David, Carol Carrado, and Dan Sichel. 2018. “The Rise of Cloud Computing: Minding Your P’s, Q’s, and K’s.” NBER Working Paper 25188.

Chen, Yongmin, and Scott J. Savage. 2011. “The Effects of Competition on the Price for Cable Modem Internet Access.” *Review of Economics and Statistics* 93 (1): 201–17.

Clark, David. 2018. *Designing an Internet.* Cambridge, MA: MIT Press.

Connolly, Michelle, and James E. Prieger. 2013. “A Basic Analysis of Entry and Exit in the US Broadband Market, 2005–2008.” *Review of Network Economics* 12 (3): 229–70.

Coyle, Diane, and David Nguyen. 2018. “Cloud Computing and National Accounting.” Economic Statistics Centre of Excellence Discussion Paper 2018-2019.

DeStafano, Timothy, Richard Kneller, and Jonathan Timmis. 2018. “Broadband Infrastructure, ICT Use, and Firm Performance: Evidence for UK Firms.” *Journal of Economic Behavior and Organization* 155: 110–39.

Ewens, Michael, Ramana Nanda, and Matthew Rhodes-Kropf. 2019. "Cost of Experimentation and the Evolution of Venture Capital." Harvard Business School Working Paper 15-070.

Federal Communications Commission. 2013. *Measuring Broadband America—2014*. Washington, DC: Federal Communications Commission.

Federal Communications Commission. 2018. *Measuring Fixed Broadband-Eighth Report*. https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-fixed-broadband-eighth-report. Accessed March, 2020.

Feldstein, Martin. 2017. "Underestimating the Real Growth of GDP, Personal Income, and Productivity." *Journal of Economic Perspectives* 31 (2): 145–64.

Goldfarb, Avi, and Catherine Tucker. 2019. "Digital Economics." *Journal of Economic Literature* 57 (1): 3–43.

Greenstein, Shane. 2010. "Glimmers and Signs of Innovative Health in the Commercial Internet." *Journal of Telecommunication and High Technology Law* 8 (1): 25–78.

Greenstein, Shane. 2015. *How the Internet Became Commercial: Innovation, Privatization, and the Birth of a New Network*. Princeton, NJ: Princeton University Press.

Greenstein, Shane, Chris Forman, and Avi Goldfarb. 2018. "How Geography Shapes—and is Shaped by—the Internet." In *The New Oxford Handbook of Economic Geography*, edited by Gordon L. Clark, Maryann P. Feldman, Meric S. Gertler, and Dariusz Wojcik, 269–85. Oxford, UK: Oxford University Press.

Greenstein, Shane, Martin Peitz, and Tomasso Valleti. 2016. "Net Neutrality: A Fastlane to Understanding the Tradeoffs." *Journal of Economic Perspectives* 30 (2): 127–50.

Greenstein, Shane, and Ryan C. McDevitt. 2011. "The Broadband Bonus: Estimating Broadband Internet's Economic Value." *Telecommunications Policy* 35 (7): 617–32.

Greenstein, Shane, and Michael Norris. 2015. "Streaming Over Broadband: Why Doesn't My Netflix Work?" Harvard Business School Case 616-007.

Groshen, Erica L., Brian C. Moyer, Ana M. Aizcorbe, Ralph Bradley, and David M. Friedman. 2017. "How Government Statistics Adjust for Potential Biases from Quality Change and New Goods in an Age of Digital Technologies: A View from the Trenches." *Journal of Economic Perspectives* 31 (2): 187–210.

Huston, Geoff. 2017. *The Rise and Rise of Content Distribution Networks*. San Diego, CA: Center for Applied Internet Data Analysis.

Internet and Television Association. 2018. "Wi-Fi: How Broadband Households Experience the Internet." *Internet and Television Association*, April 6. https://www.ncta.com/whats-new/wi-fi-how-broadband-households-experience-the-Internet.

Knieps, Günter, and Johannes M. Bauer. 2016. "The Industrial Organization of the Internet." In *Handbook on the Economics of the Internet*, edited by Johannes M. Bauer and Michael Latzer, 23–54. Cheltenham, UK: Edward Elgar Publishing.

McManus, Brian, Aviv Nevo, Zachary Nolan, and Jonathan W. Williams. 2018. "Steering Incentives and Bundling Practices in the Telecommunications Industry." Networks, Electronic Commerce, and Telecommunications Institute Working Paper 18-12.

Mueller, Milton. 2004. *Ruling the Root: Internet Governance and the Taming of Cyberspace*. Cambridge, MA: MIT Press.

Nakamura, Leonard, Jon Samuels, and Rachel Soloveichik. 2016. "Valuing 'Free Media' in GDP: An Experimental Approach." Federal Reserve Board of Philadelphia Working Paper 16-24.

Nevo, Aviv, John L. Turner, and Jonathan W. Williams. 2016. "Usage Based Pricing and Demand for Residential Broadband." *Econometrica* 84 (2): 411–43

Norton, William B. 2014. *The Internet Peering Playbook: Connecting to the Core of the Internet*. Palo Alto, CA: DrPeering Press.

Nuechterlein, Jonathan, and Phillip Weiser. 2005. *Digital Crossroads: Telecommunications Law and Economics in the Internet Age*. Cambridge, MA MIT Press.

Perrin, Andrew. 2019. "Digital Gap Between Rural and Non-rural Persists." *Pew Research Center*, May 31. https://www.pewresearch.org/fact-tank/2019/05/31/digital-gap-between-rural-and-nonrural-america-persists/.

Pew Research Center. 2019. *Internet/Broadband Fact Sheet*. Washington, DC: Pew Research Center. https://www.pewresearch.org/internet/fact-sheet/internet-broadband/.

Pew Research Center. 2019. *Mobile Fact Sheet*. Washington, DC: Pew Research Center. https://www.pewresearch.org/internet/factsheet/mobile/.

Rogerson, William P. 2018. "Economic Theories of Harm Raised by the Proposed Comcast/TWC

Transaction." Chapter 18 in *The Antitrust Revolution,* 7th ed., edited by John E. Kwoka, Jr. and Lawrence J. White. Oxford, UK: Oxford University Press.

Rosston, Gregory L., and Scott J. Wallsten. 2019. "Increasing Low-Income Broadband Adoption through Private Incentives." Unpublished. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3431346.

Russell, Andrew. 2014. *Open Standards and the Digital Age: History, Ideology, and Networks.* Cambridge, UK: Cambridge University Press.

Ryan, Camille, and Jamie M. Lewis. 2017. *Computer and Internet Use in the United States: 2015.* Washington, DC: US Census Bureau.

Seamans, Robert C. 2012. "Fighting City Hall: Entry Deterrence and Technology Deployment in the Cable TV Industry." *Management Science* 58 (3): 461–75

Simcoe, Timothy. 2012. "Standard Setting Committees: Consensus Governance for Shared Technology Platforms." *American Economic Review* 102 (1): 305–36.

Simcoe, Timothy, and Jeremy Watson. 2019. "Forking, Fragmentation, and Splintering." *Strategy Science* 4 (4): 283–97.

Syverson, Chad. 2017. "Challenges to Mismeasurement Explanations for the US Productivity Slowdown." *Journal of Economic Perspectives* 31 (2): 165–86.

US Bureau of Labor Statistics. 2020. *Consumer Price Index for All Urban Consumers. US City Average. Internet Services and Electronic Information Providers.* https//data.bls.gov/PDQWeb/cu.

Wallsten, Scott, and Colleen Mallahan. 2013. "Residential Broadband Competition in the United States." In *The Economics of Digitization,* edited by Avi Goldfarb, Shane Greenstein, and Catherine Tucker, 131–56. Cheltenham, UK: Edward Elgar Publishing.

Wang, Jin, and Kristina Steffenson McElheran. 2017. "Economies before Scale: Survival and Performance of Young Plants in the Age of Cloud Computing." Rotman School of Management Working Paper 3112901.

Zhuo, Ran, Bradley Huffaker, KC Claffy, and Shane Greenstein. 2019. "The Impact of the General Data Protection Regulation on Internet Interconnection." NBER Working Paper 26481.

Zittrain, Jonathan. 2008. *The Future of the Internet and How to Stop it.* New Haven, CT: Yale University Press.

Journal of Economic Perspectives—Volume 34, Number 2—Spring 2020—Pages 215–236

The Economics of Tipping

Ofer H. Azar

A tip is a payment from a customer to a service provider, but tips differ from other economic transactions in significant ways. The level of the tip is discretionary. It differs between customers and between occasions. The norm to tip exists in some countries and not in others. Where tipping exists, it occurs in some service occupations (like taxi drivers) but not in others (like bus drivers).

Restaurants and bars are the industries where tips are by far the most prominent. In the US food industry alone, Azar (2011) estimates that annual tips reach $47 billion. The Treasury Inspector for Tax Administration (2018) estimated a total of $44 billion of individual tip income in 2006 in the United States (including unreported tips, and not only in the food industry). These estimates make it clear that tip income in the food industry is the major source of US tip income, which is also the reason that much of the literature on tipping, including this article, tends to focus on restaurant tipping. The number of workers affected by tips is also very high. In the US, for example, there are over 2.6 million servers and 610,000 bartenders (as reported at https://www.bls.gov/ooh), occupations in which tips are a significant source of income. In many US states, tipped employees have a lower minimum wage—as low as $2.13 per hour—resulting in wage income being very low and making tips a very high part of their income. Wessels (1997) reports that the income share that comes from tips is 58 percent for servers in full-course restaurants

■ *Ofer H. Azar is a Professor in the Department of Business Administration, Guilford Glazer Faculty of Business and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel, and a Visiting Researcher in the Laboratory of Economic Behavior of the Center of Psycho-Economic Research, Povolzhsky Institute of Administration named after P.A. Stolypin-branch of RANEPA, Saratov, Russia. His email address is azar@som.bgu.ac.il.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.2.215.

and 61 percent for those in counter restaurants—and these are underestimated numbers because tips are often underreported.

For many years, tipping has received some attention from psychologists, who wanted to understand the psychological motivations for tipping, but little attention from economists. In recent years, however, tipping has attracted more interest in economics (for a detailed literature review, see Azar 2007a). The first section of this paper offers a brief historical and contextual review of tipping. The next section discusses some of the main reasons for tipping. It explains why an explanation for tipping based on rational forward-looking consumers is not supported by the evidence. Instead, tipping is better explained as a result of psychological and social motivations of consumers who obey a social norm. For tipping to encourage better service, tips should increase in service quality, and so the next step is to examine the evidence about the relationship between service quality and tips. The discussion then turns to the question of how tips often represent rents to servers and how some restaurant owners try in various ways to disrupt the social norm of tipping and instead use service charges to appropriate those rents—sometimes with the intention of redistributing them to untipped workers. An accusation sometimes leveled against tipping is that it can facilitate racial discrimination and sexual harassment, and so the evidence for these claims is outlined. The conclusion offers an overall perspective on tipping, rooted in the preceding discussion, and suggests some directions for future research.

History and Context

There is no agreement on just how the practice of tipping started. Hemenway (1993) suggests that tipping dates as far back as the Roman era and is probably even older. Segrave (1998) claims that tipping may have begun in the late Middle Ages when a master or lord of the manor could give a little extra money to a servant or laborer, whether from appreciation of a good deed or from compassion. Brenner (2001) attributes the tipping origins to 16th century England, where brass urns with the inscription "To Insure Promptitude" were placed first in coffeehouses and later in local pubs. People tipped in advance in order to get good service by putting money in these urns. Indeed, Schein, Jablonski, and Wohlfahrt (1984) and Brenner (2001) suggest that "tip" comes from the first three letters of "To Insure Promptitude," but others suggest different stories. Hemenway (1993), for example, argues that "tip" may come from stipend, a version of the Latin "stips."

It seems that Europe exported the practice of tipping to the United States, when high-income Americans who traveled in Europe in the 19th century started tipping upon their return to the United States, to show that they had been abroad and were familiar with the European customs (Schein, Jablonski, and Wohlfahrt 1984). By 1895, the average tip in European restaurants was 5 percent of the bill, while in the United States a common tip was 10 percent. Segrave (1998) estimates that during the early 1910s, five million US workers—more than 10 percent of the labor force—had tip-taking occupations. The large extent of tipping gave some tipped employees relatively high income, and employers both in Europe and the United States sometimes tried to take these economic rents from the workers either by taking the tips, or by charging employees for the right to work and earn tips (Scott 1916; Segrave 1998; Azar 2004a).

By the early 20th century, even though the tipping custom had only just arrived in the United States, there were already attempts to abolish it. Some saw tipping as creating a servants' class, part of a society where the tippers looked down upon the service providers. Gunton (1896) called tipping offensively un-American, because it was contrary to the spirit of American life of working for wages rather than fawning for favors. Some states passed laws against tipping, starting with Washington in 1909, but these laws were repealed after several years.

Over the years, the percentage tipped in the United States has gradually risen. The 10 percent tipping norm in restaurants in the late 19th century stayed for several decades (Hathaway 1928; Post 1937), but eventually increased to 15 percent (Post 1984). In her etiquette manual, Post (1997) writes, "It wasn't long ago that 15 percent of the bill, excluding tax, was considered a generous tip in elegant restaurants. Now the figure is moving toward 20 percent for excellent service. In ordinary family-style restaurants 15 percent is still the norm." Today, some travel guides refer to 15–25 percent as a tipping standard in restaurants (for example, https://www.lonelyplanet.com/news/2016/12/07/how-much-to-tip/). A similar pattern of increasing tip percentages is observed in taxi tipping, starting with 10 percent early in the 20th century (Hathaway 1928), rising in mid-century to 15 percent (Post 1984), and then by the end of the 20th century reaching 20 percent in large cities (Post 1997).

Today, tipping norms differ around the globe. Tourist guidebooks often provide advice about the tipping norms in the country (for example, Star 1988). In Europe, where tipping originated and was common already hundreds of years ago, today tipping is generally less common and in much smaller magnitudes than in the United States. In many European restaurants, tipping takes the form of rounding up the restaurant bill a little, not adding 15–20 percent to the bill. Along with restaurant servers and taxicab drivers, some of the professions where tipping is at least relatively common include food delivery people, bartenders, and hair salon workers.

Main Reasons for Tipping

It is extremely difficult to explain tipping behavior from the perspective of a purely self-interested consumer. After all, a tip is usually given after the service has been provided. One can make a theoretical argument that if the self-interested consumer is in a repeated relationship with a seller, for a good or service where qualities controlled by the seller matter, then the consumer could have an economic motivation to tip. However, this motivation for tipping is irrelevant for customers without intentions to return to the same service provider. In a big city, for example, you are unlikely to meet the same taxi driver in the future—and even if you did,

Table 1
Reasons Given for Tipping

Reason for tipping	United States	Israel
1. Avoid feeling guilty	60.2%	13.3%
2. Avoid embarrassment	44.1%	23.2%
3. Tipping being a social norm	84.7%	58.1%
4. Show gratitude	67.8%	68.9%
5. Waiters depend on tips	66.9%	32.4%
6. Get poor future service if I don't tip	13.6%	2.5%
7. The waiter may yell at me if I don't tip	4.2%	0.0%
Total number of reasons	3.42	1.98
Number of observations	118	241

Source: Based on Azar (2010).

the driver is unlikely to recognize you and adjust the quality of service to your tip a couple of years ago. People also tip in restaurants or hotels when they travel and do not plan to go back to the same places. This already suggests that for many tippers (those who are one-time customers), future service is not a reason for tipping. In what follows, I will explain why even for repeat customers, future-service considerations are not likely to be a significant reason for tipping. Instead, psychological and social motivations seem more plausible reasons for tipping. Then, I also discuss how the motivations of tippers, workers, and restaurants could combine to create a self-reinforcing social norm that sustains tipping behavior.

The Lack of Support for Future-Minded Tipping

One way to gain insight into why people tip is to ask them. Azar (2010) conducted a survey in the United States and Israel, asking people to mark the reasons (possibly more than one) why they tip, giving seven possible reasons. Table 1 reports some results. Future-service considerations or the risk of being yelled at do not seem to be an important reason in either the United States or Israel. In both countries, the most common responses are that tipping is a social norm and to show gratitude. However, US respondents (who gave more reasons on average compared to Israelis) also often endorsed the reasoning that waiters depend on tips and that they want to avoid feeling guilty and avoid embarrassment as reasons why they tip.

Additional evidence that future service does not motivate tipping is provided by Azar (2007b). Azar builds a theoretical model showing that if future service is a reason for tipping, the sensitivity of tips to service quality should be higher for repeat customers than for one-time customers. Azar then analyzes empirical data from restaurants and does not find evidence for such difference in tipping behavior, suggesting that future service is not a reason for tipping even by repeat customers.

An alternative approach to assess the importance of future interactions as a potential cause for tipping is to look at the occupations where tipping is common. Lynn (2016) uses a survey to study characteristics of tipped versus non-tipped service occupations and finds that higher likelihood of encountering the same service provider on multiple service occasions and occupations with higher frequency of customer patronage are not more likely than other occupations to be tipped. As a concrete example, it seems unlikely that anyone thinks about the chance of meeting the same cab driver for a future trip when calculating a tip. This again suggests that future-service motivations are not an important reason for tipping. In a similar spirit, Lynn (2018) uses an online survey to study the motivations for tipping in various occupations. He finds a wide array of occupations where future-service motives are plausible, but tipping is rare. Some of these rarely tipped occupations where a consumer might plausibly expect to see the same person on a future visit were dental hygienists, veterinarians, car mechanics, appliance delivery and installation men, and restaurant workers who hand you the food when you pick up takeout orders. This again suggests that future-service motivations are not an important motivation for tipping.

Because future-service motivations do not seem to be a significant reason for tipping, we can conclude that the vast majority of tipping is not motivated by economic considerations of a selfish and emotionless consumer, but rather by psychological and social motivations, and in particular the first five reasons in Table 1. Indeed, tipping as the social norm was the most popular reason given for tipping in the US and the second most popular in Israel. Moreover, the reasons of avoiding feeling guilty and avoiding embarrassment, which are frequently endorsed, are also related to tipping being a social norm; people do not feel guilty or embarrassed not to tip workers where there is no social norm to tip them (like flight attendants).

A different approach to understand the motivations for tipping is provided by Azar (2004b), who combines a model of tipping with the historical evidence. In the model, a customer chooses the size of a tip (measured as a percentage of the bill) to maximize utility in a framework where the choice of a tip can provide higher utility from following a social norm or feeling good personally about the tip (say, from feeling generous), but a higher tip also costs more.[1] In this model, the social norm is endogenous and evolves over time: the norm in each period is the average tip in the previous period. The model allows customers not to have any positive feelings from tipping beyond the desire to obey the social norm, but the analysis shows that if tips remain unchanged or increase over time, this means that at least for some customers, such positive feelings from tipping exist. Combined with the historical evidence discussed earlier about tip percentages in the United States increasing

[1] Specifically, the utility function in Azar (2004b) is

$$u(g;\, n_t, \theta) = d(g - n_t) + \theta p(g) - bg,$$

where g is the tip in percentage of the bill, n_t is the tipping norm at time t, and p is an increasing function that captures the psychological utility from tipping that results from feeling generous, impressing others, etc. The value of θ captures the degree of positive feelings that the customer gains from tipping, where $\theta \geq 0$ and the population of customers may be heterogeneous in the value of θ. The function d captures the disutility from social disapproval and the desire to conform to the social norm, and it is assumed to be maximized at $d(0)$, and b is the bill size. The customer chooses g to maximize utility (subject to $g \geq 0$), taking as given the values of n_t and θ.

over time, the model suggests that the US pattern during the last century is that people tipped on average a little above the social norm, because it gave them positive feelings of being generous, impressing others, showing their gratitude, and so on. Because the average tipper tipped above the norm, over time the tipping norm increased. Then people started tipping above this increased norm, the norm further increased, and so on. This process can explain why the percentage tipped in US restaurants roughly doubled over the last century. The model can be applied also to the evolution of social norms other than tipping, and it suggests that when a norm is costly to follow and people do not derive benefits from following it other than avoiding social disapproval, the norm erodes over time.

Tipping as a Welfare-Increasing Social Norm

Why are social norms, like tipping, created? One view in economics is that social norms are created because they increase welfare. In this spirit, Arrow (1971) argued:

> I want, however, to conclude by calling attention to a less visible form of social action: norms of social behavior, including ethical and moral codes. I suggest as one possible interpretation that they are reactions of society to compensate for market failures…There is a whole set of customs and norms which might be similarly interpreted as agreements to improve the efficiency of the economic system (in the broad sense of satisfaction of individual values) by providing commodities to which the price system is inapplicable.

In this spirit, Azar (2005a) examines whether tipping is a norm that increases social welfare using a model of tipping in which a customer and a server interact. The server's utility is standard and includes the earnings from the tip and the cost of effort, both depending on the service quality, which the server chooses optimally. The customer's utility function is less standard, as it includes psychological utility and a social norm, along with the cost of the tip itself. However, there is no assumption that following the social norm exactly will lead to the highest utility, which allows for cases in which motivations like generosity and empathy for the server result in the psychological utility being maximized at a tip level that is above the social norm.[2] Azar finds, in line with the view expressed by Arrow (1971), that tipping can indeed improve

[2] Specifically, the customer's utility function in Azar (2005a) is

$$u(s, t, b) = -b(1 + t) + f(t - n(s)) + G(s, b),$$

where s is service quality, t is the tip percentage, b is the bill size, and $n(s)$ is the social norm function about how much should be tipped for each service quality. The first term of $-b(1 + t)$ captures the cost of dining and leaving the tip. The last term of $G(s, b)$ is the utility from the dining experience, which depends on service quality and on the quantity and quality of food purchased, which are captured by the bill size. The function f captures the psychological utility from feelings and depends on the difference between the amount tipped (t) and the norm about how much should be tipped for the service quality received.

service quality and social welfare, and that these improvements will be larger when the sensitivity of tips to service quality according to the social norm is higher.

Which Occupations Are Tipped?

One more approach to thinking whether the purpose of tipping is to improve social welfare is to look across occupations. Azar (2005b) uses a list of 37 occupations, some of which are tipped whereas others are not. Using several sources on tipping practices, Azar builds a measure of tipping prevalence in each occupation. Using questionnaires, the occupations were rated on several dimensions: the closeness established between the customer and the worker during the service; the worker's income not including tips; the income of the typical customer of this service occupation; the relative importance of the personal versus professional components of service; and the ability of the customer to evaluate the professional component of service. The latter two are used to create a measure of the customer's ability to monitor the worker. The results suggest that tipping prevalence is negatively correlated with the worker's income (before tips) and positively with the customer's income, so tipping seems to serve as a redistributive norm: it happens more in occupations where it transfers money from high-income individuals to low-income ones. Tipping prevalence is positively affected by the closeness established during the service between the worker and the customer. In terms of economic efficiency, if tipping was created in the occupations where it has the highest ability to lower monitoring costs by shifting them from the employer to the customer, we should see a positive relationship between the customer's monitoring ability and tipping prevalence; however, this relationship is negative (though not statistically significant) in the data. However, if we extend the notion of social welfare to include psychological utility, then possibly tipping is created where it improves social welfare the most, because tipping seems to be more prevalent in occupations where consumers feel empathy and compassion for the workers and want to show gratitude for good service.

Lynn (2016) takes a similar approach, using a more extensive list of 122 service occupations and a different set of questions. He concludes that in the United States, customers are more likely to tip occupations with these characteristics: workers are less happy than customers during the service encounter; the workers' income, skill, and required judgment are low; the workers deliver customized service; and customers can more easily evaluate workers' performance than managers can. The last item implies that tipping is more common where it yields higher economic efficiency.

Customer Interests Regarding Tipping

A social norm in favor of tipping might be based on the interaction of psychological, social, and economic motivations between customers and sellers. For customers, tipping appeals to several different motivations, as discussed above, and consequently most customers prefer tipping to the main alternatives. For example, Azar (2010) asked in his survey if customers would prefer tipping or an automatically added service charge. Roughly one-fifth of the US and Israeli respondents expressed no preference between the two. However, those who did express a preference were overwhelmingly in favor of tipping in both countries (tipping was over four times more popular on average), despite the modest level of service charge proposed as the alternative (15 percent of the bill in the United States and 12 percent of the bill in Israel).

Why do customers prefer tipping? The Azar (2010) survey did not ask directly, but the responses to other questions provide some hints. Over two-thirds of the respondents, in both the US and Israel, endorsed this reason for tipping: "By tipping I can show the waiter my gratitude for his service." In addition, respondents reported a high sensitivity of their tips to service quality, suggesting that they want to be able to control what they pay for the service, according to the actual service they receive. When asked for the reasons they tip more for better service, the majority of respondents reported that it is fairer that tips will depend on service, and that they want to show their gratitude in proportion to how grateful they actually are. These responses can explain why tipping is preferred to service charges that do not depend on service quality.

Restaurant Interests Regarding Tipping

Servers often earn much more with tipping than they would in a regime in which tipping is replaced by service charges or higher prices that go to the restaurant, which would only pay them wages. What about the interests of the restaurant owner? I will discuss at a later point why restaurants may not want to support tipping. But do restaurants also have reasons to prefer tipping to service charges or higher prices? The answer is "yes."

First, because most customers prefer tipping and restaurants operate in a competitive environment, it makes sense (at least for most restaurants) to choose the policy that most customers prefer.

Second, higher prices in lieu of tipping may result in customers perceiving the restaurant as more expensive and less attractive, even if the tips plus the lower prices are equal to the increased prices. This can be the result of psychological biases in the perception of prices, which puts more emphasis on the menu prices than on the additional costs (such as tips).

Third, the mechanism of tipping, in which a customer receives service and observes its quality and then has discretion regarding how much to tip, can potentially serve as customer monitoring that improves service and social welfare (Azar 2005a). Imagine that it is too costly for the owner to supervise all the dimensions of service quality for certain workers, like their degree of friendliness (and then also pay more to workers who provide better service). It is possible that tipping provides an incentive to improve service, in a way that saves money for the restaurant and makes both the worker and the customer better-off.

Fourth, if a restaurant replaces tipping with higher prices and then pays its servers what the average server used to earn with tips, it benefits those who earned lower tips at the expense of those who earned higher tips. The latter servers may quit and go to work at a competing restaurant that uses tipping, where they will earn more. Presumably the servers who had previously received higher tips are the ones the customers like more (and therefore they tip them more). Thus, the restaurant's move to higher prices results in self-selection of the poorer servers to continue working and the better servers to leave.

Finally, we should remember that we are usually talking not about a new restaurant, but about an existing restaurant, and tipping is the status quo. Replacing tipping with higher prices at an existing establishment is a change. Changes create various costs, they may have unforeseen consequences, and firms often do not like changes without clear benefits. Changing a restaurant with tipping to a restaurant with higher prices requires printing new menus. It risks upsetting repeat customers. Weekend shifts that are popular among the servers with tipping due to higher tip income will become unattractive due to more work for the same pay under fixed wages, possibly making it hard to get servers for these busy shifts.

Taking these motivations together, a restaurant has an incentive to facilitate tipping by making it easy and salient for tips to be paid. One way we see this happening is that US restaurants and taxicabs commonly offer the bill with a space for a tip to be inserted for those who pay with credit cards, and sometimes they even calculate and show the customer how much a tip of 15, 20, or 25 percent would be. The observation that all the relevant parties—the customers, the servers, and the restaurants—have reasons to retain tipping can explain why it is still such a widespread phenomenon and a social norm that at least in some industries and some countries survives and even flourishes and becomes more significant.

The Relationship between Tipping and Service Quality

Does tipping actually provide incentives to the workers in a manner that improves service quality? The answer depends on the sensitivity of tips to service quality, and several studies have tried to assess this sensitivity empirically. A common methodology used for this purpose is to give customers as they exit the restaurant a questionnaire about their dining experience, which includes questions about service quality and usually some additional control variables. The bill and tip amounts can be collected either from the customers or from the restaurant. Regressions in which the tip is the dependent variable and the service quality (as rated by the customer) is one of the independent variables then suggest to what extent tips are associated with service quality.

How Is Service Quality Associated with Tipping?

The results about the relationship between service and tips from customer surveys are mixed. On one hand, studies that examine whether a certain behavior of servers affects tips often find a significant effect. For example, Lynn, Le, and Sherwyn (1998) found that customers tipped significantly more when they were touched by the server than when they were not touched, and Lynn and Mynier (1993) found that squatting during the initial visit to the table increases tips.

On the other hand, when considering the relationship between service quality and tips directly, the sensitivity of tips to service quality seems to be low. In a meta-analysis, Lynn and McCall (2000) found a statistically significant and positive relationship between tips and service evaluations, but the effect of service on tips was small and accounted for less than 2 percent of the variability in tip percentage (although it approached 5 percent of the variability in studies that used customer ratings of service on multi-item scales, which may be more reliable).

Several similar studies since that review have also found that the sensitivity of tips to service quality is positive, but modest. Conlin, Lynn, and O'Donoghue (2003) collected survey data from 1,393 customers of 39 restaurants in Houston, Texas. They averaged the customers' answers to four measures of service quality (appearance, friendliness, speed of service, and attentiveness)[3] and combined these to create a single measure on a 1–5 scale. They find that an increase of 1 point in this combined measure raises the tip by 1.46 percentage points, which is about 8 percent above the average tip of 17.56 percent. Azar (2007b) finds in a sample of 597 diners from six restaurants that repeating customers (N=169) reduce their tips by 0.09 percent of the bill for each additional point in self-reported service quality (on a 0–5 scale), whereas non-repeating customers (N=428) increase their tips by 0.90 percent of the bill for each point, representing about a 6.8 percent increase in tip income.

What Azar (2009) called the "tipping-service puzzle" arises here. The service ratings in the last two studies mentioned are generally very good, with an average score of about 4.5 out of 5. On the other hand, the low sensitivity of tips to service quality suggests that servers may be better off making less effort and giving lower service quality, because their tips will hardly be affected. How is low sensitivity of tips to service quality consistent with high service quality? One likely answer is that restaurants have many ways of trying to assure the expected high quality of service: training, monitoring, rules for server behavior, the threat of a customer complaining to the manager, and so on. A second possibility is that the server can provide some elements of better service like being friendly and smiling at low cost. Finally, servers may perceive the connection between tips and service quality to be stronger than it actually is. For example, Kwortnik, Lynn, and Ross (2009, in their Study 2) recruit 1,189 current servers from 48 US states to participate in a survey. Half of the servers indicated that service had a large or very large effect on tips, 37 percent indicated a medium effect, and only 13 percent indicated a small or very small effect.

Interestingly, when people are asked about their general tipping behavior rather than about how much they tipped on a certain occasion, their answers suggest a very high sensitivity of tips to service quality. Azar (2010) reports that on average, when service quality increases from 1 to 5 (on a 1–5 scale), people claim that they

[3] They also collected customers' ratings for the servers' "knowledge of menu." However, this rating turned out to be negatively correlated with the tips, and the authors suggest that this may be a result of knowledgeable servers bothering the customer with unimportant information or being perceived as a "snob." Therefore, this measure is not integrated with the other four measures.

increase their tips from 8.7 to 24.5 percent (in the US sample) or from 6.2 to 21.0 percent (in the Israeli sample).

How Does Tipping Affect Service Quality?

Several studies seek to assess how tipping affects service quality by comparing establishments with tipping to ones with compulsory service charges or service-inclusive prices. Kwortnik, Lynn, and Ross (2009, their Study 1) take advantage of the fact that about 40 percent of the restaurants in Miami Beach, Florida replaced voluntary tipping with automatic service charges, because many customers are international tourists unfamiliar with the US tipping norms. For 107 restaurants, the authors obtained the restaurant's Zagat service rating on a 0–30 scale and their tipping policy (voluntary tipping or a service charge, based on the policy towards tables of four or fewer customers). The authors regress the Zagat service rating on the tipping policy, controlling for food rating, décor rating, and restaurant expensiveness. The results show that service rating has a greater positive association with restaurants that practice voluntary tipping than with those having a service charge.

Lynn and Brewster (2018) analyze online ratings of restaurants that changed their policy from tipping to automatic service charges or service-inclusive pricing. They find that eliminating tipping reduced online customer ratings—more strongly when tipping is replaced with automatic service charges than when it is replaced with service-inclusive pricing, but at a level which is highly statistically significant in both cases. They also find that more expensive restaurants suffer from a smaller reduction in online customer ratings after eliminating tipping compared to cheaper restaurants. A natural interpretation of this result is that more expensive restaurants probably hire more experienced and professional servers, train them better, can afford (and need) to invest more in monitoring them, and pay them more, and therefore have more mechanisms that ensure high service quality also when the incentives provided by tipping disappear.

Lynn and Kwortnik (2015) also find a positive effect of voluntary tipping on service. They take advantage of the fact that Carnival Cruises changed its tipping policy from voluntary tipping to an automatic service charge in the early 2000s. Using reviews of 1,129 customers who rated the service on ten ships, the authors find that service ratings under tipping are significantly higher than under service charges.

We should remember that the studies that compare service quality with and without tipping may be comparing establishments that differ in other ways, too. For example, it is plausible that restaurants or cruise ships without tipping invested more in screening potential employees, training them, monitoring them, finding ways to motivate them financially, and otherwise to provide excellent service. Therefore, these studies may not be comparing tipping versus no-tipping with other factors held constant, but rather comparing tipping with a combination of no-tipping but a greater emphasis on other incentives for quality service. Given that these other incentives should improve service, the finding that nevertheless service is better with tipping reinforces the conclusion that tipping improves service compared to no-tipping. In other words, the estimated improvement in service quality due to tipping in these studies may in fact underestimate the real improvement.

Tipping and a Struggle over Rents

Those in restaurants who receive tips are relatively low-skill workers who often earn more (in wages and tips) than other workers of similar skill levels—including those who work in the kitchen of the same restaurant. One example of this pattern became public when Danny Meyer, the head of the restaurant group Union Square Hospitality Group, announced in October 2015 that he would eliminate tipping in all the group's restaurants (Steiner 2017). The initial plan was to have all 13 of the group's restaurants fully tipless by the end of 2016 (Sutton 2015). Meyer claimed that compensation for kitchen workers had increased by about 25 percent since 1985, but increased more than 300 percent for tipped employees in the same period (Ruggless 2018). This resulted in difficulties in hiring chefs and cooks, who earn much less than servers (Ferdman 2015). A cook in New York makes between $10-$12 per hour and may face the burdens both of big-city rents and repaying culinary school debts. However, restaurants find it difficult to increase cooks' wages because they have thin profit margins; the National Restaurant Association estimates that the median profit margin for mid-level establishments (having average checks of more than $25) was 4.5 percent (Ferdman 2015).

Meyer mentioned (as quoted in Sutton 2015) that "we can't have a situation where we are asking someone to pay $40,000 to go to the Culinary Institute of America to then work for $12.50 per hour, when they could work in fast food for $15." Servers at full-service New York restaurants can expect 20 percent tips in most cases, reaching a yearly income of $40,000 (representing $20 per hour) or more on average, with some of the city's top servers earning above $100,000 annually (Sutton 2015). The big gap in compensation, where cooks earned much less than servers because they were not tipped, led many cooks to switch to being servers, even though they had previously invested large amounts in culinary education.[4] Similarly, Azar (2011) cites a California restaurant owner who switched from tips to service charges and wrote: "Before starting this service charge, our servers were making, on average, over TWICE the wage of the cooks, and while servers require only a few months of training, our cooks require 2–5 years!"

Restaurant owners have tried in several ways to claw back some of these rents. In the US, one approach in states that allow it is to pay only the "tipped minimum wage," which is a lower legal hourly minimum wage that employers can pay workers

[4] Meyer, for example, mentioned on a *Freakonomics* podcast that "when I learned a statistic that for the first time in my entire career, that we had more culinary grads working in the dining room than in the kitchen, that was the moment when I said, 'That has to stop,' because they didn't go to cooking school to be servers" (Dubner 2016).

who make money from tips. If tips and the tipped minimum wage together do not sum up to the regular minimum wage, the employer must fill the gap so that the worker gets in total at least the minimum wage. The federal tipped minimum wage is only $2.13, although many states have higher tipped minimum wages.[5]

Another approach is for restaurants to switch from tipping to other ways to charge for service. Azar (2012) argued (before the recent trend of eliminating restaurant tipping) that increasing the minimum wage for tipped workers, and thus increasing the rents of servers, can lead restaurants to replace tipping with automatic service charges or with service-inclusive pricing. When a restaurant makes this change, it can pay the servers the regular minimum wage (or more, if their reservation wage is higher) and keep for itself the extra amounts (from the service charges or price increases due to service-inclusive pricing) above this wage (that is, the economic rents), which the servers kept under tipping.

The shifts to a higher tipped minimum wage in some places in recent years offers an opportunity to examine whether this indeed leads to changes in restaurant policy towards tipping. More generally, it allows us to examine to what extent a strong social norm such as tipping can be changed when one of the main players—the restaurant—finds that this social norm is unprofitable. The prediction of Azar (2012) indeed has been validated in recent years. In 2015, New York State decided to increase the tipped minimum hourly wage from $5 to $7.50 (Paarlberg and Reyes 2018), with further increases scheduled for the future. In New York City, for example, the tipped minimum hourly wage increased to $10 for employers of 11 or more workers at the start of 2019 and to smaller employers at the start of 2020 (as reported at https://labor.ny.gov/formsdocs/wp/Part146.pdf). This is far above the federal tipped minimum hourly wage, which is $2.13.

This was also the time when Danny Meyer announced that he would eliminate tipping for the 13 restaurants in the Union Square Hospitality Group (Steiner 2017), and other restaurants also announced plans to eliminating tipping and replacing it by service-including prices or automatic service charges. At about this time, Crowley and Rami (2016) mention a survey of 503 American restaurateurs who were asked about adopting a no-tipping policy: 18 percent said they have done so already, and 29 percent intended to follow suit. Only 27 percent said they are pro-tipping and will stay with tips. Another 10 percent said they were undecided, and 17 percent mentioned that they might switch to no-tipping if more of their colleagues do. The survey results reflect that initially, there was a lot of enthusiasm about Meyer's move to eliminate tipping in his restaurants. In a *Freakonomics* podcast (Dubner 2016), Meyer described how everyone was happier after the move to no-tipping in one restaurant. Due to the great publicity of the change, more customers came. This

[5] For a list of higher tipped minimum wages across states, see https://en.wikipedia.org/wiki/Tipped_wage. Among the four most populated states, which account for over one-third of the US population, only Texas uses the $2.13 federal tipped minimum hourly wage. California has no tip credit, and so the minimum wage of tipped workers is equal to the regular minimum wage and is above $10/hour; New York has a tipped minimum hourly wage of $7.50 to $10, depending on the location and employer; and Florida has a tipped minimum wage of $5.23/hour.

allowed Meyer to increase wages to back-of-the-house (kitchen) workers and to avoid lowering the servers' wages. Servers no longer received tips, but their wages rose from $5/hour to $9/hour, plus a revenue-sharing program. Job applications for the kitchen went up 270 percent, but even the server applicant pool increased by 25 percent, 100 percent, and then 215 percent in the first three months. Meyer declared in late 2015 that the move of Union Square Hospitality Group (USHG) to hospitality included (increased prices and no tipping) will not hurt the average waiter, saying: "The waiters at our restaurants, when we eliminate tipping, will make as much or more, in 75 percent of cases, than they're making right now" (as quoted in Steiner 2017).

But after the initial excitement, it became gradually understood across the restaurant industry that there is no free lunch: that is, you cannot charge customers the same amount, increase the salaries of kitchen workers, and have the servers earn the same as before. The plan to switch all Union Square Hospitality Group (USHG) restaurants to no-tipping in one year was delayed. Many former and current servers reported that their income dropped substantially after the no-tipping policy was instituted, by about $100 per week (Steiner 2017). Many servers quit.[6] As noted earlier, busy shifts at a restaurant require more effort from a server, but with a fixed hourly wage and no tips, this extra work does not offer more income. Erin Moran, USHG's chief culture officer, suggests that customer visits at the restaurants dropped, reducing servers' compensation due to the revenue-share program. She said: "We learned very quickly that this is much more challenging than we had anticipated." Under-qualified and under-trained workers were sometimes hired to replace those who quit (Steiner 2017). Starting salaries for cooks did increase from about $11/hour to $13/hour, but because overtime is now managed more carefully, overall compensation for cooks is only slightly higher than it was under the tipping system. Servers can no longer increase their income by getting better shifts or higher tips, and to get a raise, they need to apply for it with a manager and pass written tests on their knowledge of service, food, and wine (Steiner 2017).

Problems with the no-tipping policy were not limited to the 13 Union Square Hospitality Group (USHG) restaurants. Some other restaurants that tried no-tipping, including the national chain Joe's Crab Shack, eventually switched back to tipping (Crowley and Rami 2016; Steiner 2017; Sun 2018). Thus, while switching to no-tipping gives the restaurant the possibility to extract economic rents from servers, it has to be done carefully in order not to reduce servers' income beyond the level they find acceptable, not to drive down service quality, and not to upset customers. Assuming that the service charges or price increases that replace tipping are at about the same

[6] When the no-tipping "hospitality included" program was introduced at Maialino restaurant in 2016, almost the entire front-of-house staff turned over, twice. At North End Grill, a significant group of workers from both the front-of-house and back-of-house quit. A former manager at Union Square Cafe mentioned that "the majority of our strongest people have left" (Steiner 2017). Meyer also admits that some servers were not thrilled with the no-tipping policy and many of them quit (Ruggless 2018).

magnitude as the forgone tips, the customers do not perceive themselves as better off without tipping. Many customers prefer the tipping policy where they have more control over how much they pay for service, where they can feel that they are demonstrating generosity, and according to the studies mentioned earlier, they also receive better service.

If servers' economic rents continue to rise due to increases in the tipped minimum wages, restaurants will probably continue the attempts to take away these rents by adopting no-tipping policies, even if it takes some time to figure out how to make the no-tipping policies successful. New York, for example, has considered abolishing the tip credit entirely (Albany Bureau 2018), which means that tipped workers will receive full minimum wages, which in New York will soon be $15 per hour in many places. Earning tips of about 20 percent in addition to a higher minimum wage implies that servers will earn far more than others with similar occupational characteristics, which does not seem to be a stable equilibrium.

The analysis here suggests that those who are worried about the pay disparity within restaurants and who want to improve the situation of restaurant workers who are not tipped should not advocate to eliminate tipping. Instead, they should advocate that the restaurant be allowed to impose "tip pooling," in which the restaurant would allocate some of the tips from servers to the kitchen workers.

Tipping, Discrimination, and Sexual Harassment

The voluntary aspect of tipping raises the possibility that tips will be distributed in a discriminatory manner. It raises a further question about whether different groups disburse tips in the same way—and if they don't, might this result in the different groups receiving different levels of service. I present some evidence bearing on these questions, and then discuss whether tipping is linked to sexual harassment.

Discrimination Affecting Those Who Receive Tips

Several studies analyze how tipping differs by the ethnicity of the tipper and the tipped worker. Ayres, Vars, and Zakariya (2005) study tipping in New Haven taxis, using 1,066 surveys of tipping data from twelve drivers, consisting of six black drivers, four white drivers, and two "other minority" drivers (one self-reported as "Arab" and the other as "Asian (Indian)"). Table 2 presents some of their results: the rows show the race or ethnicity of the passengers, while the columns show the 12 taxicab drivers divided into white, black, and "other."

The bottom row of Table 2 shows that whites in this group receive higher tips. A similar pattern is observed when counting the rate at which drivers are not tipped at all: white drivers are stiffed by 15.7 percent of customers, black drivers by 28.3 percent, and the others by 36.4 percent. The final column shows that white passengers tipped more. Again, stiffing rates by passengers' race gives a similar picture, with 10.6 percent stiffing by white passengers, 39.2 percent by black

Table 2
Tips Received by New Haven Cab Drivers
(Number of observations in parentheses)

	Cab driver race			
Passenger race	White	Black	Other	Average tip from passengers of this race
White	26.7% (224)	17.9% (222)	13.2% (34)	21.6%
Black	11.0% (112)	7.4% (176)	13.1% (28)	9.2%
Hispanic	17.5% (57)	7.1% (63)	11.3% (17)	12.0%
Asian	16.1% (42)	18.1% (36)	12.3% (17)	16.2%
Other	14.8% (6)	11.0% (5)	1.8% (3)	10.7%
Average tip for this group of drivers	20.3%	12.6%	12.4%	

Source: Derived from Ayres, Vars, and Zakariya (2005).

passengers, 34.3 percent by Hispanic passengers, 15.8 percent by Asian passengers, and 35.7 percent by other passengers. Several interesting patterns emerge in the intersections of the table. White passengers tipped white drivers 1.49 times more than black drivers. Perhaps surprisingly, black passengers also tipped white drivers 1.49 times more than black drivers; this may suggest that the higher tips to white drivers are not due to prejudice but due to a higher perceived service quality. Hispanic passengers seem to discriminate the most against black drivers, tipping white drivers 2.46 times more than black drivers. However, Asian passengers in this sample tip black drivers a little more than white drivers: 18.1 percent to black drivers versus 16.1 percent to white drivers.

Lynn et al. (2008) study ethnic differences in restaurant tipping, collecting data from 140 lunch customers of a southern US restaurant. Both white and black customers on average tipped white servers more than black servers. Interestingly, the form of this effect differed by perceived service quality. For non-perfect service quality (rating less than 5 on 1–5 scale), average tips for black servers (16.6 percent) were similar to tips for white servers (16.8 percent). However, for perfect service quality, white servers earned much higher tips averaging 23.4 percent, whereas black servers still received 16.6 percent.

While discrimination in tipping is an unwanted consequence, calls to abolish tipping because of it go too far. A tip in a place like a US restaurant resembles in some ways a gift in a wedding: both are voluntary in principle, but virtually compulsory in practice in the sense that almost everyone gives them, and yet the giver does retain discretion about how much to spend.[7] We will not abolish wedding gifts even if we find racial differences in the amounts received. Similarly, tipping should not be abolished due to racial differences. Instead, attempts to eliminate discrimination in tipping should use education and discussion. In addition, in some situations and in particular in restaurants, pooling the tips of the various workers and distributing them equally can solve the inequality in pay (although it also reduces the incentives of the servers to provide excellent service).

Differences in Tipping across Groups

The New Haven taxicab study suggests that the tipper's race is associated with the size of the tip, and this finding is consistent with other studies. Brewster and Nowak (2018) present the aggregate results of 11 previous studies that were published between 2003 and 2017 and used different methodologies: server records, exit surveys of diners, a telephone survey, and online surveys. In all 11 studies, white customers tipped on average more than black customers. In the aggregate data across all studies, the mean tip percent of black customers (N = 1,156) is 14 percent, while the mean tip of white customers (N = 4,003) is 18 percent.

Given the robust ethnic differences in tipping behavior, a natural question is whether servers discriminate against those who are expected to be poor tippers. Brewster and Rusche (2012) address this question using survey data derived from a community sample of servers (N = 200). The data indicate that servers do perceive tables with white customers to be better tippers than black tables. In addition, servers reported that white tables are easier to wait on than comparable black tables. Survey answers about self-reported server discrimination are reported in Table 3.

In these self-reported data, some servers do give poorer service to minorities, and it seems plausible that this behavior is motivated by prejudice more than by economic incentives related to tipping. Moreover, a server with pure economic motivations is not motivated by the average tip, but by the sensitivity of tips on the margin to service quality, and this sensitivity is not necessarily lower for minority groups. In fact, Ayers, Vars, and Zakariya (2005) find that the rates at which minority groups provide no tip at all to taxicab drivers are much higher. If this finding also applies to restaurant tipping, it suggests that the sensitivity of minority groups to service quality could be higher than that of other groups, giving incentives to make more effort when serving them.

Tipping and Sexual Harassment

One concern about tipping is that it may make customers feel entitled to practice sexual harassment and other behaviors that are unpleasant to the servers, especially female servers (Einhorn and Abrams 2018). However, most of the sexual

[7] In some places, such as Israel, the similarity is even higher because almost everyone gives a monetary gift. Daum (2016) reports that only 1 percent of her respondents believed that gift-receivers would enjoy getting a personal gift at their wedding, with the other options being cash or check.

Table 3
A Survey of Servers on Discrimination

	Answers			
Questions	Never	Sometimes	Often	Always
How often does the quality of service that you provide vary according to the customers' race?	61.5%	31.8%	4.6%	2.1%
How often do you give customers poor service because you think they will not tip well?	50.8%	42.1%	5.6%	1.5%
How often do you observe poor treatment of black customers by coworkers?	47.1%	38.7%	8.9%	5.2%
How often do you think your coworkers give customers poorer service than other customers based on the customers' race?	17.9%	54.4%	9.7%	1.5%

Source: Brewster and Rusche (2012).
Note: Rows may not sum to 100 percent because "don't know" category is left out. Sample size is N = 200.

harassment cases in restaurants are not by customers, and when such harassment does occur, it does not seem to have any close linkage to tipping.

For example, Johnson and Madera (2018) study sexual harassment by following 76 female US college students working in food and beverage service, mainly restaurants. These 76 students documented any sexual harassing behaviors they encountered over three months. The most frequent behaviors selected were when someone at work "told suggestive, sexual stories" (reported by 49 percent), "made offensive remarks" (46 percent), "made crude sexual remarks" (45 percent), "made sexist comments" (42 percent), and "attempted to discuss sex" (33 percent)." Of the 226 incidents recorded (an average of about one incident per month per worker), 112 of the incidents involved coworkers, 29 involved a manager, and 85 involved customers. However, there is no obvious reason why tipping is to blame, and it is not clear that limitations or abolition of tipping would diminish such behaviors.

A different national-level survey finds what seems to be much worse sexual harassment. McColl (2017) surveyed 300 hospitality workers, 90 percent of them women, and 89 percent had experienced sexual harassment at work. Of the respondents, 87 percent reported sexist remarks, 85 percent reported comments on their body, and 84 percent reported sexual innuendo. In addition, 73 percent said they had experienced unwanted sexual advances, 69 percent reported inappropriate touching, and 19 percent said they had been sexually assaulted. However, this survey was done in Australia, and as a travel guide notes, "generally, Australians and New Zealanders say tipping is not only unnecessary but also a practice to be avoided …" (Rivera 2019). Clearly, sexual harassment can be a real and pervasive problem in the restaurant and hospitality industry, but it's not at all clear that tipping is the culprit.

Conclusion

Tipping is for obvious reasons a subject of controversy. A social norm that dictates a tip as a percentage of the bill doesn't seem to make intuitive sense: opening a $100 bottle of wine does not require quadruple the effort of opening a $25 bottle of wine, and bringing a $40 steak is not four times as tiring as bringing a $10 pasta. Critics have argued that tipping enforces archaic and undesirable social distinctions and that it facilitates discrimination in who holds what jobs in a restaurant, as well as encouraging sexual harassment (Ferdman 2016).[8] A number of writers have expressed a negative attitude towards tipping or even called to abolish it (for example, Porter 2013; Keane 2014; S.K. 2015; Grimminck 2015; Dixon 2016).

The discussion and evidence in this paper suggests a more nuanced view. Perhaps some restaurants, under a certain kind of management and with a certain kind of reputation, could benefit by replacing tipping with a fixed service charge or with higher menu prices. Indeed, some restaurant owners strongly believe that replacing tipping with a fixed service charge, distributed among both servers and kitchen help, led to a better quality of food, improved team spirit, and ultimately resulted in a higher level of professionalism and better service—even if some previously high-tipped servers experienced a fall in income (for example, see Porter 2013).

However, at many other restaurants, the social norm of tipping has been reasonably robust. Many customers prefer the control of choosing a tip and have a positive feeling that they are showing generosity. Servers earn more as a result and find that busy shifts where they have to work harder are rewarded with higher income. Service quality seems modestly higher. It seems likely that restaurant owners, tempted by the tip rents being earned by servers, will continue to experiment with ways of recapturing some of those rents—either to redistribute to kitchen workers or to raise profits. Perhaps the self-reinforcing social norm of tipping will be toppled eventually in the United States, but with more than a century of history, it seems unlikely to go quickly.

The literature on the economics of tipping is relatively small. Some additional economic and business issues related to tipping and some future research directions are discussed in Azar (2003, 2007c, 2011). For example, one issue is the possibility that a tipped employee will behave in a manner that is harmful to an employer in order to increase tips (for example, by giving the customers free desserts or drinks).

[8] Ferdman (2016) interviews Saru Jayaraman, who is the co-founder and co-director of the Restaurant Opportunities Center United (ROC United) and director of the Food Labor Research Center at the University of California, Berkeley. Ferdman mentions that "Jayaraman has been working to end the modern tipping system for years." Jayaraman describes evils of tipping (such as encouraging sexual harassment and servers preferring the interests of the customer to those of the restaurant), she relates tipping to slavery, and she mentions that tipping has racial implications because workers of color at fine-dining restaurants get jobs that are non-tipped instead of becoming servers and bartenders. Eventually, despite these evils, Jayaraman advocates for keeping tipping, but just moving the tip credit towards the minimum wage, saying "we're fine with that—we're not trying to get rid of tipping entirely. We just want the remaining forty three states to follow in California and the other six states' footsteps, eliminating the two-tiered system and requiring employers to pay the full wage to all workers, with tips on top of that."

The recent move of some restaurants from tipping to alternatives such as service charges or higher prices raises interesting questions and provides opportunities for theoretical, empirical, and experimental work. The possibility of tip pooling, either between tipped employees of the same occupation (like servers) or between tipped and non-tipped employees (like servers and kitchen workers) also affects incentives and has economic implications. Tipping norms are different around the globe, which should allow for some interesting comparisons and research possibilities. One can also imagine the possibility of future shifts in the social norm of tipping, perhaps with tips spreading to other occupations or tips becoming more sensitive to both service quality and to effort more generally.

■ *I am grateful to the editors Gordon Henson, Enrico Moretti, Timothy Taylor, and Heidi Williams for their valuable comments and to a grant from the Government of the Russian Federation (research project "Cognitive-behavioral and cross-cultural foundations of economic policy," Contract 14.W03.31.0027) for financial support. I have no relevant or material financial interests that relate to the research described in this paper.*

References

Arrow, Kenneth. 1971. "Political and Economics Evaluation of Social Effects and Externalities." In *Frontiers of Quantitative Economics*, edited by Michael D. Intriligator, 3–25. Amsterdam: North-Holland.

Ayres, Ian, Fredrick E. Vars, and Nasser Zakariya. 2005. "To Insure Prejudice: Racial Disparities in Taxicab Tipping." *Yale Law Journal* 114: 1613–74.

Azar, Ofer H. 2003. "The Implications of Tipping for Economics and Management." *International Journal of Social Economics* 30 (10): 1084–94.

Azar, Ofer H. 2004a. "The History of Tipping—From Sixteenth-Century England to United States in the 1910s." *Journal of Socio-Economics* 33 (6): 745–64.

Azar, Ofer H. 2004b. "What Sustains Social Norms and How They Evolve?: The Case of Tipping." *Journal of Economic Behavior and Organization* 54 (1): 49–64.

Azar, Ofer H. 2005a. "The Social Norm of Tipping: Does It Improve Social Welfare?" *Journal of Economics* 85 (2): 141–73.

Azar, Ofer H. 2005b. "Who Do We Tip and Why? An Empirical Investigation." *Applied Economics* 37 (16): 1871–79.

Azar, Ofer H. 2007a. "The Social Norm of Tipping: A Review." *Journal of Applied Social Psychology* 37 (2): 380–402.

Azar, Ofer H. 2007b. "Do People Tip Strategically, to Improve Future Service? Theory and Evidence." *Canadian Journal of Economics* 40 (2): 515–27.

Azar, Ofer H. 2007c. "Why Pay Extra? Tipping and the Importance of Social Norms and Feelings in Economic Theory." *Journal of Socio-Economics* 36 (2): 250–65.

Azar, Ofer H. 2009. "Incentives and Service Quality in the Restaurant Industry: The Tipping-Service Puzzle." *Applied Economics* 41 (15): 1917–27.

Azar, Ofer H. 2010. "Tipping Motivations and Behavior in the U.S. and Israel." *Journal of Applied Social Psychology* 40 (2): 421–57.

Azar, Ofer H. 2011. "Business Strategy and the Social Norm of Tipping." *Journal of Economic Psychology* 32 (3): 515–25.

Azar, Ofer H. 2012. "The Effect of the Minimum Wage for Tipped Workers on Firm Strategy, Employees and Social Welfare." *Labour Economics* 19 (5): 748–55.

Brenner, Mark L. 2001. *Tipping for Success: Secrets for How to Get in and Get Great Service.* Sherman Oaks: Brenmark House.

Brewster, Zachary W., and Sarah Nell Rusche. 2012. "Quantitative Evidence of the Continuing Significance of Race: Tableside Racism in Full-Service Restaurants." *Journal of Black Studies* 43 (4): 359–84.

Brewster, Zachary W., and Gerald Roman Nowak III. 2018. "Racial Prejudices, Racialized Workplaces, and Restaurant Servers' Hyperbolic Perceptions of Black-White Tipping Differences." *Cornell Hospitality Quarterly* 60 (2): 159–73.

Conlin, Michael, Michael Lynn, and Ted O'Donoghue. 2003. "The Norm of Restaurant Tipping." *Journal of Economic Behavior and Organization* 52 (3): 297–321.

Crowley, Chris, and Trupti Rami. 2016. "Despite Setbacks, a New Survey Says More Restaurants around the Country Will Go Tip-Free." *New York Grub Street*, June 3. http://www.grubstreet.com/2016/06/danny-meyer-tipping-survey.html.

Daum, Liora. 2016. "Courtesy versus Efficiency: Personal Gifts and Monetary Gifts—Preferences and Norms in Israeli Society." MA thesis. Ben-Gurion University of the Negev.

Dixon, Vince. 2016. "The Case against Tipping in America." *Eater.* https://www.eater.com/a/case-against-tipping.

Dubner, Stephen J. 2016. "The No-Tipping Point." *Freakonomics Radio*, March 9. http://freakonomics.com/podcast/danny-meyer/.

Einhorn, Catrin, and Rachel Abrams. 2018. "The Tipping Equation." *New York Times*, March 12. https://www.nytimes.com/interactive/2018/03/11/business/tipping-sexual-harassment.html.

Ferdman, Roberto A. 2015. "The Crippling Problem Restaurant-Goers Haven't Noticed but Chefs Are Freaking Out About." *Washington Post*, August 12. https://www.washingtonpost.com/news/wonk/wp/2015/08/12/the-crippling-problem-people-who-eat-at-restaurants-havent-noticed-but-chefs-are-freaking-out-about/?utm_term=.4c4535e52f7e.

Ferdman, Roberto A. 2016. "I Dare You to Read This and Still Feel Good about Tipping." *Washington Post Wonkblog*, February 18. https://www.washingtonpost.com/news/wonk/wp/2016/02/18/i-dare-you-to-read-this-and-still-feel-ok-about-tipping-in-the-united-states/?utm_term=.b6fcef3e13ec.

Grimminck, Robert. 2015. "10 Reasons to Abolish the Tipping System in Restaurants." *TopTenz*, March 25. http://www.toptenz.net/10-reasons-to-abolish-the-tipping-system-in-restaurants.php.

Gunton, George. 1896. "Economic Effects of Tipping." In *Gunton's Magazine*, 13–19. New York: Political Science Publishing Co.

Hathaway, Helen. 1928. *Manners: American Etiquette.* New York: E.P. Dutton.

Hemenway, David. 1993. Prices & Choices: Microeconomic Vignettes. Lanham: University Press of America.

Johnson, Stefanie K., and Juan M. Madera. 2018. "Sexual Harassment Is Pervasive in the Restaurant Industry. Here's What Needs to Change." *Harvard Business Review*, January 18. https://hbr.org/2018/01/sexual-harassment-is-pervasive-in-the-restaurant-industry-heres-what-needs-to-change.

Keane, Tom. 2014. "Let's Ban Tips: Tradition, Not Logic, Keeps Antiquated Practice in Place." *Boston Globe*, December 1. https://www.bostonglobe.com/opinion/2014/12/01/let-ban-tips/EdQ28Eg1SuwqR9p7XaG23J/story.html.

Kwortnik, Robert J. Jr., W. Michael Lynn, and William T. Ross Jr. 2009. "Buyer Monitoring: A Means to Insure Personalized Service." *Journal of Marketing Research* 46 (5): 573–83.

Lynn, Michael. 2016. "Why Are We More Likely to Tip Some Service Occupations Than Others? Theory, Evidence, and Implications." *Journal of Economic Psychology* 54: 134–50.

Lynn, Michael. 2018. "How Motivations for Tipping Vary with Occupational Differences in Descriptive Tipping Norms." *Journal of Behavioral and Experimental Economics* 77: 1–10.

Lynn, Michael, and Zachary W. Brewster. 2018. "A Within-Restaurant Analysis of Changes in Customer Satisfaction Following the Introduction of Service Inclusive Pricing or Automatic Service Charges." *International Journal of Hospitality Management* 70: 9–15.

Lynn, Michael, and Robert J. Kwortnik. 2015. "The Effects of Tipping Policies on Customer Satisfaction: A Test from the Cruise Industry." *International Journal of Hospitality Management* 51: 15–18.

Lynn, Michael, Joseph-Mykal Le, and David S. Sherwyn. 1998. "Reach Out and Touch Your Customers." *Cornell Hotel and Restaurant Administration Quarterly* 39 (3): 60–65.

Lynn, Michael, and Michael McCall. 2000. "Gratitude and Gratuity: A Meta-Analysis of Research on the Service-Tipping Relationship." *Journal of Socio-Economics* 29: 203–14.

Lynn, Michael, and Kirby Mynier. 1993. "Effect of Server Posture on Restaurant Tipping." *Journal of Applied Social Psychology* 23 (8): 678–85.

Lynn, Michael, Michael C. Sturman, Christie Ganley, Elizabeth Adams, Matthew Douglas, and Jessica McNeal. 2008. "Consumer Racial Discrimination in Tipping: A Replication and Extension." *Journal of Applied Social Psychology* 38 (4): 1045–60.

McColl, Gina. 2017. "Sleaze, Assault and Even Threats of Rape Are All on the Menu for Young Women Who Work in Hospitality—So Why Does No One Stop it?" *Sydney Morning Herald*. https://www.smh.com.au/interactive/2017/harassment-in-hospitality/.

Paarlberg, Michael A., and Teófilo L. Reyes. 2018. *New York's Experience after the Tipped Minimum Wage Increase*. New York and Washington, DC: Restaurant Opportunities Centers United and Institute for Policy Studies.

Porter, Jay. 2013. "What Happens When You Abolish Tipping." *Slate*, August 14. https://slate.com/human-interest/2013/08/tipless-restaurants-the-linkerys-owner-explains-why-abolishing-tipping-made-service-better.html.

Post, Emily. 1937. *Etiquette: The Blue Book of Social Usage*. New York: Funk and Wagnalls.

Post, Elizabeth L. 1984. *Emily Post's Etiquette*. 14th ed. New York: Harper and Row.

Post, Peggy. 1997. *Emily Post's Etiquette*. 16th ed. New York: William Morrow and Co.

Rivera, Larry. 2019. "Is Tipping Mandatory in Australia?" *TripSavvy*, August 20. https://www.tripsavvy.com/is-tipping-mandatory-in-australia-1464392.

Ruggless, Ron. 2018. "Danny Meyer Talks Tipping, Leadership and Trust." *Restaurant Hospitality*, February 5. https://www.restaurant-hospitality.com/people/danny-meyer-talks-tipping-leadership-and-trust.

Schein, John E., Edwin F. Jablonski, and Barbara R. Wohlfahrt. 1984. *The Art of Tipping: Customs & Controversies*. Wausau: Tippers International.

Scott, William R. 1916. *The Itching Palm: A Study of the Habit of Tipping in America*. Philadelphia: Penn Publishing Company.

Segrave, Kerry. 1998. *Tipping: An American Social History of Gratuities*. Jefferson, NC: McFarland and Company.

S.K. 2015. "Service Compris: The Case against Tipping." *The Economist*, October 26. https://www.economist.com/gulliver/2015/10/26/the-case-against-tipping.

Star, Nancy. 1988. *International Guide to Tipping*. New York: Berkeley Publishing Group.

Steiner, Keenan. 2017. "As Danny Meyer Continues to Eliminate Tips, His Staff Struggles to Adapt." *New York Grub Street*, October 19. http://www.grubstreet.com/2017/10/danny-meyer-no-tips-staff-expectations.html.

Sun, Mengqi. 2018. "After Trying No Tipping Policies, Some New York Restaurants Reverse Course: Some Staff and Customers Have Been Resistant to Gratuity-Included Menus." *Wall Street Journal*, September 29. https://www.wsj.com/articles/after-trying-no-tipping-policies-some-new-york-restaurants-reverse-course-1538233200.

Sutton, Ryan. 2015. "Danny Meyer Is Eliminating All Tipping at His Restaurants: Nothing Will Ever Be the Same." *Eater New York*, October 14. https://ny.eater.com/2015/10/14/9517747/danny-meyer-no-tipping-restaurants.

Treasury Inspector for Tax Administration. 2018. *Billions in Tip-Related Tax Noncompliance Are Not Fully Addressed and Tip Agreements Are Generally Not Enforced*. Washington, DC: Department of the Treasury.

Vaughn, Natasha. 2018. "NY May End the Tipped Wage. What to Know." *Democrat and Chronicle*, March 15. https://www.democratandchronicle.com/story/news/politics/albany/2018/03/15/should-new-york-ended-tipped-wage-servers-what-you-should-know/419623002/.

Wessels, Walter J. 1997. "Minimum Wages and Tipped Servers." *Economic Inquiry* 35 (2): 334–49.

Journal of Economic Perspectives—Volume 34, Number 2—Spring 2020—Pages 237–244

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by email at taylort@macalester.edu, or c/o *Journal of Economic Perspectives*, Macalester College, 1600 Grand Ave., St. Paul, MN 55105.

Potpourri

The National Science Foundation has published its biennial report *The State of U.S. Science and Engineering 2020* (January 2020, https://ncses.nsf.gov/pubs/nsb20201). "Among individual countries, the United States was the largest R&D performer in 2017, followed by China, whose R&D spending now exceeds that of the EU. Together, the United States (25%) and China (23%) accounted for nearly half of the estimated global R&D total in 2017. Japan (8%) and Germany (6%) are next, followed by South Korea (4%). France, India, the United Kingdom, Russia, Brazil, Taiwan, Italy, Canada, Spain, Turkey, and Australia account for about 1%–3% each of the global total." "Foreign-born workers—ranging from long-term U.S. residents with strong roots in the United States to more recent immigrants—account for 30% of workers in

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota. He blogs at http://conversableeconomist.blogspot.com.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.34.2.237.

S&E occupations. The number and proportion of the S&E workforce that are foreign born has grown. In many of the broad S&E occupational categories, the higher the degree level, the greater the proportion of the workforce who are foreign born. More than one-half of doctorate holders in engineering and in computer science and mathematics occupations are foreign born." "A substantial percentage of Americans also think science makes life change too fast (49%)."

The Joint Center for Housing Studies of Harvard University discusses "America's Rental Housing 2020" (January 2020, https://www.jchs.harvard.edu/sites/default/files/Harvard_JCHS_Americas_Rental_Housing_2020.pdf). "[R]entership rates for all age groups under 65 are still historically high. . . . [T]he number of high-income renters continued to climb. . . . In fact, households with real incomes of at least $75,000 accounted for over three-quarters of the growth in renters (3.2 million) from 2010 to 2018, while the number earning less than $30,000 fell by nearly 1 million. . . . This represents a sharp reversal of trends in the 2000s, when low-income households drove 93 percent of renter growth and the number of high-income households declined by 160,000. This shift has significantly altered the profile of the typical renter household. When rentership rates hit bottom in 2004 during the homeownership boom, 18 percent of renters earned $75,000 or more and 42 percent earned less than $30,000. By 2018, this disparity had narrowed considerably, with high-income households accounting for 23 percent of renters and low-income households for 38 percent. . . . [F]rom the homeownership peak in 2004 to 2018, the number of married couples with children that owned homes fell by 2.7 million, while the number renting rose by 680,000. . . . Ownership of rental housing shifted noticeably between 2001 and 2015, with institutional owners such as LLCs, LLPs, and REITs accounting for a growing share of the stock. Meanwhile, individual ownership fell across rental properties of all sizes, but especially among buildings with 5–24 units."

S. Çelik, G. Demirtaş and M. Isaksson review "Corporate Bond Market Trends, Emerging Risks and Monetary Policy" (OECD Capital Market Series, February 2020, http://www.oecd.org/corporate/ca/Corporate-Bond-Market-Trends-Emerging-Risks-Monetary-Policy.pdf). "By the end of 2019, the global outstanding stock of non-financial corporate bonds reached an all-time high of USD 13.5 trillion in real terms. . . . [I]n comparison with previous credit cycles, today's stock of outstanding corporate bonds has lower overall rating quality, higher payback requirements, longer maturities and inferior investor protection." "In every year since 2010, around 20% of the total amount of all bond issues has been non-investment grade and in 2019 the portion reached 25%. This is the longest period since 1980 that the portion of non-investment grade issuance has remained so high, indicating that default rates in a future downturn are likely to be higher than in previous credit cycles." "As of December 2019, non-financial companies worldwide need to repay or refinance an unprecedented USD 1.3 trillion within on year, USD 2.9 trillion within 2 years and USD 4.4 trillion within 3 years. The amount due within 3 years represents a record 32.4% of the total outstanding amount."

John P. McGowan and Ed Nosal wrote "How Did the Fed Funds Market Change When Excess Reserves Were Abundant?" (*Economic Policy Review*, Federal Reserve

Bank of New York, March 2020, https://www.newyorkfed.org/medialibrary/media/research/epr/2020/epr_2020_vol26no1.pdf). "Prior to the 2007 financial crisis, trading in the fed funds market was dominated by banks. Banks managed the balances—or reserves—of their Federal Reserve accounts by buying these balances from, or selling them to, each other. . . . The amount of excess reserves in the banking system—total reserves minus total required reserves—was very small and banks actively traded fed funds in order to keep their reserves close to the required amount. . . . In the post-crisis abundant excess reserves environment, the amount of bank-to-bank lending in the fed funds market was quite small. Instead, lending in the fed funds market became dominated by the Federal Home Loan Banks (FHLBs). These institutions have accounts at the Fed, which allows them to trade fed funds, but they are ineligible to receive IOER [interest on excess reserves] compensation. . . . As a result, FHLB balances behave in a manner similar to that of other non-reserve account holders at the Fed, such as the foreign repo pool or the Treasury General Account. . . . The rationale for borrowing in the federal funds market also changed during the abundant reserves period. Such borrowing became dominated by foreign banking organizations (FBOs), which used the market as a source of funding and arbitrage profit. . . . As such, the fed funds market was not typically used to make adjustments to banks' reserve holdings vis-à-vis their required holdings since reserve holdings far exceeded what was required."

Symposia

Shelly Lundberg has edited *Women in Economics*, which contains an introduction and 18 short essays (VoxEU.org, March 2020, free registration required, https://voxeu.org/content/women-economics). From Lundberg's "Introduction," "The representation of women in academic economics in the United States grew substantially during the 1970s and 1980s in the wake of a wave of feminist activism throughout the academy. . . . [T]hat progress stalled in the ensuing decades, with the share of female assistant professors and PhD students remaining roughly constant since the mid-2000s. . . . [I]n Europe. . .as in the US, the representation of women falls from about 40% at the entry level to 22% among full professors and that a cohort effect alone cannot explain this change. . . . [E]conomics research papers written by women appear to be held to higher standards in the publishing process than papers written by men. As in several other professions (medicine, real estate, law), there appears to be a quality/quantity trade-off, with female economists producing less output of higher quality than equivalent men. . . . [I]n the findings of their study of gender differences in the collaborative networks of economists, . . . women work with a smaller network of distinct co-authors than men and tend to collaborate repeatedly with the same co-authors and their co-authors' collaborators, constructing a tighter network. Since larger networks are associated with higher levels of research output, these patterns may disadvantage women." Some chapters in the book offer a condensed version of arguments made at greater length in the “Symposium on Women in Economics” in the Winter 2019 issue of this journal (https://www.aeaweb.org/issues/538).

The *BIS Quarterly Review* offers a seven-paper symposium about global exchange rate markets (Bank for International Settlements, December 2019, https://www.bis.org/publ/qtrpdf/r_qt1912.htm). From the overview essay by Philip Wooldridge, “FX and OTC derivatives markets through the lens of the Triennial Survey,” “FX and OTC derivatives markets saw a marked pickup in trading between the 2016 and 2019 surveys. Following a dip in 2016, FX trading returned to its long-term upward trend, rising to $6.6 trillion per day in April 2019. . . . The trading of short-term instruments grew faster than that of long-term instruments. . . . [T]he trading of FX swaps, which is concentrated in maturities of less than a week, rose from $2.4 trillion in April 2016 to $3.2 trillion in April 2019 and accounted for most of the overall increase in FX trading. . . . In FX markets, London, New York, Singapore and Hong Kong SAR increased their collective share of global trading to 75% in April 2019, up from 71% in 2016 and 65% in 2010. Trading in OTC interest rate derivatives markets was also increasingly concentrated in a few financial centres, especially London. . . . [I]t is more cost-effective to centralise counterparty and credit relationships, or technical and legal infrastructures, in a handful of hubs than to spread them across many countries. The faster pace of trading also increased the advantages of locating traders’ IT systems physically close to those of the platforms on which they trade.”

The *Journal of the American Planning Association* offers two viewpoint articles advocating the abolition of single-family zoning, along with seven short commentaries and then two rejoinders (2020, 86:1, https://www.tandfonline.com/toc/rjpa20/86/1). Here’s a metaphor from Michael Manville, Paavo Monkkonen and Michael Lens in their essay, “It’s Time to End Single-Family Zoning”: “Suppose that, for your wellbeing, you need regular access to only a small amount of expensive medicine. One day you go to the pharmacy and learn the government has implemented a new rationing system strictly limiting the number of sales that can occur in small doses. Because many people, like you, only need small doses, the new rule results in few small doses being available. Plenty of medicine is available—you can see it over the counter—but the pharmacist can only sell it in large quantities. So you are stuck. If you want your medicine, you must buy more than you need, at a price higher than you can afford. This new rationing system is also strictly enforced. Not only must you buy in large quantities, but you cannot divide up your ration afterward and sell your extra doses to others who might need and value them. Most people, we suspect, would consider such a rationing system unjust and inefficient. It would force a large number of people to spend and consume more than they otherwise would, subsidize the smaller number of people who want and can afford large doses, and keep some people from getting medicine at all. Fortunately, the United States does not allocate medicine in this bizarre manner. But it does ration urban land this way.”

Pathways magazine has devoted a special issue of 13 short essays about the “Millenial Dilemma,” concerning those in the age bracket from 23 to 38 (Stanford Center on Poverty and Inequality 2019, https://inequality.stanford.edu/publications/pathway/state-union-2019). Michael Hout argues, “American men and women born since

1980—the millennials—have been less upwardly mobile than previous generations of Americans. The growth of white-collar and professional employment resulted in relatively high occupational status for the parents of millennials. Because that transition raised parents' status, it set a higher target for millennials to hit." Florencia Torche and Amy L. Johnson posit, "Millennials with no more than a high school diploma have much lower earnings in early adulthood than prior generations." Darrick Hamilton and Christopher Famighetti state, "Young millennials have lower rates of homeownership than Generation X, baby boomers, and the Silent Generation at comparable ages. We have to reach back to a generation born nearly a century ago—the Greatest Generation—to find homeownership rates lower than those found today among millennials. The racial gap in young-adult homeownership is larger for millennials than for any generation in the past century." Kim A. Weeden points out, "The gender segregation of occupations is less pronounced among millennials than among any other generation in recent U.S. history. By contrast, millennials are experiencing just as much racial and ethnic occupational segregation as prior generations, even though millennials are less tolerant of overt expressions of racism."

Future of Children contains a seven-paper symposium on the theme of "Universal Approaches to Promoting Healthy Development" (Princeton-Brookings, Spring 2019, https://futureofchildren.princeton.edu/sites/futureofchildren/files/media/foc_vol_29_no_1_0.pdf.) From the introductory essay by Deborah Daro, Kenneth A. Dodge, and Ron Haskins, "If we saw children in a canoe heading for a waterfall, we wouldn't be content to wait at the bottom and mend their wounds after they crash; instead, we would climb to the top of the falls and try to stop them from going over the edge in the first place. Similarly, we must begin earlier in children's lives and come equipped to identify family needs and offer assistance in a timely fashion before problems surface. . . . [T]he programs described in this issue suggest that building the infrastructure to support a universal approach involves three important steps. The first is a mechanism that allows all families in a community to be assessed and to receive advice from qualified professionals about how to ensure the healthy development of their children. Second, when giving advice, these professionals must recommend specific prevention or treatment services that are available and accessible in the local community—and, often, help parents gain access to and pay for the services. Third, programs need a way to track families' developmental and other issues, the services they seek and receive, and the results of those services. . . . [W]e are guided by growing evidence that well-crafted and carefully implemented prevention strategies can significantly improve parental capacity and child safety while also enhancing child development."

About Economists

Jessie Romero has an interview with Janice Eberly in *Econ Focus* (Federal Reserve Bank of Richmond, Fourth Quarter 2019, pp. 22–26, https://www.richmondfed.org/publications/research/econ_focus/2019/q4/interview). "We're familiar with investments in physical capital, by which I mean property, plant, and equipment — the things most people would recognize as capital. That's tangible capital. But today we also have intangible capital—the investments you can't touch, such as software and intellectual property. You can expand the definition to include things like worker skills that are specific to the firm; when a firm invests in its employees, it's also developing its capital in some broad sense. . . . These types of investments are increasingly important: Intangible capital is the fastest-growing part of investment. It also seems to be playing a greater role in the success of firms. Not only is intangible capital a larger and larger share of investment overall, but it's also especially important for the firms that end up being the leading firms in their industries. Amazon's business is built on intangible capital; Walmart's logistics technology is all intangible capital. . . . Historically, we thought technological change was embodied in tangible capital. . . . So the question is whether physical capital is embodying technological change in the way that it used to. Is the technological change actually in the intangible capital? . . . Just like job growth has shifted toward the service jobs you can't send overseas, investment has shifted toward the industries where you can't offshore the capital and away from the durable goods and manufacturing industries."

Tyler Cowen conducts one of his *Conversations with Tyler*: "Daron Acemoglu on the Struggle Between State and Society" (Medium.com, December 4, 2019, https://medium.com/conversations-with-tyler/tyler-cowen-daron-acemoglu-history-economics-db440127ba6a, both audio and transcript available). "[I]t turns out that there's one surefire predictor of when a country democratizes — it's economic crisis. Dictatorships don't go often because they decide, well, citizens should rule themselves. They collapse, and they collapse more likely in the midst of severe economic recessions. . . . So when a country democratizes, for another three or four years, it takes time for it to get out of the crisis. Then it starts a much faster growth process. It's not going to make Nigeria turn into Switzerland, but a country that democratizes adds about 20 to 25 percent more to its GDP per capita. . . . [O]ne of the most important mechanisms for that seems to be that when you democratize, you tax more, so the taxation, the budgets go up. And you spend more, especially on education and health, so the health of the population improves. Child mortality is one of the things that improves very fast. Primary and secondary enrollment improves a little bit more slowly, but it improves very steadily."

Stephen E. Landsburg has written *The Essential Milton Friedman* (Fraser Institute, 2019, https://www.essentialscholars.org/friedman). It's a free e-book, 73 pages long, with highly readable nonspecialist overview of many of Friedman's most prominent ideas. "Milton Friedman was one of the most influential economists of all time. He revolutionized the way economists think about consumption, about money, about stabilization policy, and about unemployment. . . . In several cases, Friedman's methods inspired the creation of entire new subfields including the economic analysis of law, the quantitative approach to economic history, the economics of crime and punishment, the economics of family relationships, and the economic approach to finance. . . . And he influenced policymakers. In the United States, he helped to end the military draft, to broaden educational choice, and to change the regulatory climate. Worldwide, almost all central banks now follow policies that are grounded in Friedman's insights and recommendations. . ." The Fraser Institute has also published e-books *The Essential Adam Smith* by James Otteson (2018, https://www.essentialscholars.org/smith) and *The Essential Hayek*, by Donald J. Boudreaux (2014, https://www.essentialscholars.org/hayek).

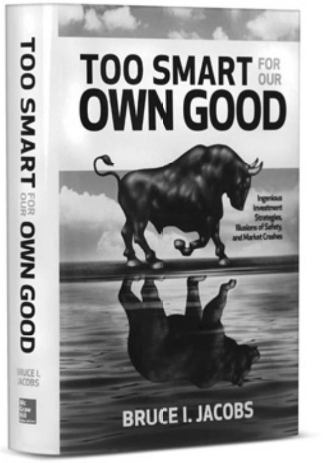
Discussion Starters

Jeff Cockrell inquires "Does America Have an Antitrust Problem?" (*Chicago Booth Review*, Winter 2019/2020, https://review.chicagobooth.edu/economics/2019/article/does-america-have-antitrust-problem). "'Concentration isn't a good barometer of the extent of competition in the market,' says Chicago Booth's Chad Syverson. 'It's not just a noisy barometer; we don't even know what direction the needle is pointing. There are cases where, clearly, things happen in a market to make it more competitive, and concentration goes up.' Consider a market in which customers face high switching costs, or barriers from moving from one seller to another. The mobile-phone market prior to the advent of number portability, or the right to take your phone number with you when switching carriers, was a good example: the inconvenience of transitioning to a new phone number imposed a high cost on those who wanted to leave their provider for a new one. If switching costs come down, the market becomes more competitive, but at the same time, many inferior suppliers will lose market share or go out of business as their customers abandon them for better options, resulting in higher concentration. . . . 'The issue is, what's the criterion that you use to declare something as noncompetitive in a way that's easily comparable across industries?' Syverson says. Some have suggested that size—measured by total asset value, for instance—could be such a criterion, but 'I think that is just taking a hatchet to a birthday cake,' he says. 'It's ridiculously blunt.'" This article reflects themes in the "Symposium on Markups" and "Symposium on Issues in Antitrust" from the Summer 2019 issue of this journal (https://www.aeaweb.org/issues/559).

In "No More Californias," Kyle Mangum writes, "As American mobility declines, some wonder if we've lost our pioneer spirit. A closer look at the data suggests that the situation is less dire—and more complicated—than it at first appears" (*Economic Insights*: Federal Reserve Bank of Philadelphia, Winter 2020, pp. 8–13, https://www.philadelphiafed.org/-/media/research-and-data/publications/economic-insights/2020/q1/eiq120-go-west.pdf). "There are two perspectives on this question. One is that the expansion of population across the continent was simply a phase in the life cycle of American development. . . . Maybe there was nothing uniquely American about high mobility (besides, perhaps, open land) and no reason to desire it now. The wagons reached the coast, and there were no more Californias to settle. In this case, there is no problem for policy to fix. The second perspective is that population change is unduly restricted by policy failures that create congestion in desirable, productive places. Regulations that make it hard to build new homes increase costs and prevent cities, especially those offering high incomes or many amenities, from adding new residents. Suboptimal urban planning could lead cities to be overly congested and below capacity. . . . These two perspectives are not mutually exclusive, and the reality likely combines the two. The regional transition is mostly complete (subject to the caveat that there is always potential for new shocks), and the new trend in population growth is in the expansion of existing cities (especially those away from the coasts) across various regions. This should assuage the fears raised by the interregional migration decline, and there is really no clear role for policy here anyway. The real question is whether this natural new phase of population growth is producing the optimal distribution of population across cities, especially across cities within each region."

For those on the lookout for some good news about what economists have done, Michael Kremer, Jonathan D. Levin, and Christopher M. Snyder offer "Advance Market Commitments: Insights from Theory and Experience" (February 2020, NBER Working Paper 26775). "Ten years ago, donors committed $1.5 billion to a pilot Advance Market Commitment (AMC) to help purchase pneumococcal vaccine for low-income countries. The AMC aimed to encourage the development of such vaccines, ensure distribution to children in low-income countries, and pilot the AMC mechanism for possible future use. Three vaccines have been developed and more than 150 million children immunized, saving an estimated 700,000 lives. This paper reviews the economic logic behind AMCs, the experience with the pilot, and key issues for future AMCs." Back in the Fall 2002 issue of this journal, Kremer was already making arguments for advance market commitments in "Pharmaceuticals and the Developing World" (https://www.aeaweb.org/articles?id=10.1257/089533002320950984).

Too Smart for Our Own Good


TOO SMART FOR OUR OWN GOOD
Ingenious Investment Strategies, Illusions of Safety, and Market Crashes
BRUCE I. JACOBS
Mc Graw Hill Education

Ingenious Investment Strategies, Illusions of Safety, and Market Crashes

Bruce I. Jacobs

One of today's leading financial thinkers, Bruce I. Jacobs, examines recent financial crises—including the 1987 stock market crash, the 1998 collapse of the hedge fund Long-Term Capital Management, the 2007–2008 credit crisis, and the European debt crisis—and reveals the common threads that explain these market disruptions. In each case, investors in search of safety were drawn to novel strategies that were intended to reduce risk but actually magnified it—and blew up markets. Until we manage risk in responsible ways, major crises will always be just around the bend. *Too Smart for Our Own Good* is a big step toward smarter investing—and a better financial future for everyone.

"Bruce Jacobs explains when a crash is likely: It's when the economy is strong and risks appear to be low. Buy this book today and be forewarned."
—Elroy Dimson, Professor of Finance, Cambridge Business School

"Bruce Jacobs's insightful analyses of financial crises will alert readers to how some financial instruments and strategies can mask investment risk and lead to excessive leverage. Investors and financial institutions would do well to heed the warnings in this book."
—Frank J. Fabozzi, Professor of Finance, EDHEC Business School, and Editor,
The Journal of Portfolio Management

"Bruce Jacobs takes a close look at financial blowups over four decades and finds a common element: risk management and investment strategies that appear benign at the micro level but pose dire systemic risks at the macro level."
—Greg Feldberg, Director of Research, US Financial Crisis Inquiry Commission

"*Too Smart for Our Own Good* is a remarkable combination of decades of hands-on wisdom from a great investor with astute analytical insight born of detailed research—on a topic that is vital not only to the world of finance, but also to the world at large."
—Geoffrey Garrett, Dean, The Wharton School

"The increasing frequency of market crashes is a clarion call for a thorough investigation of the causes of market fragility. *Too Smart for Our Own Good* offers a critical analysis that is of paramount importance for all of us."
—Michael Gibbons, Deputy Dean, and Professor of Finance, The Wharton School

About the Author

Bruce I. Jacobs is co-founder, co-chief investment officer, and co-director of research at Jacobs Levy Equity Management. He is co-author, with Ken Levy, of *Equity Management: The Art and Science of Modern Quantitative Investing*. Jacobs serves on the Advisory Boards of the *Journal of Portfolio Management* and *Journal of Financial Data Science*, and has served on the *Financial Analysts Journal* Advisory Council. He holds a Ph.D. in finance from The Wharton School.

Visit: www.mhprofessional.com
ISBN: 9781260440546
Available in print, ebook, and audiobook formats

RESERVE BANK
Applications open for the AEA Summer Program
News
acer

ASK • ADVISE • SHARE

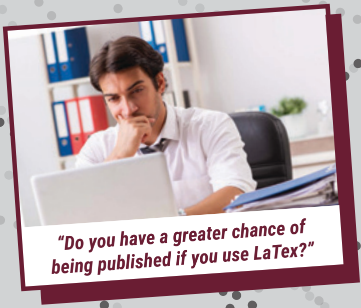
Join the conversation!

New features include:

- *Multiple-mode posting option*
- *Polls • Inline images*

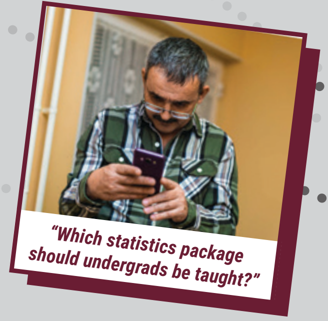
"Have low interest rates led to excessive risk taking?"

"What are the best sources for job market advice?"



"Do you have a greater chance of being published if you use LaTex?"

"What are some dissertation-worthy topics in market design?"



"Which statistics package should undergrads be taught?"

www.aeaweb.org/econspark

Participate in a community where economists at all career levels learn from and advise each other in a professional and supportive online environment.

AMERICAN ECONOMIC ASSOCIATION

ECON spark

AEA Discussion Forum

The American Economic Association presents...

RESEARCH HIGHLIGHTS

A Convenient Way to Monitor Key Economics Research and Emerging Topics Published in AEA Journals

Can trade policy drive economic development?

As income inequality rises, are living standards diverging?

Do medical bankruptcy laws affect health insurance rates?

The American Economic Association now publishes a weekly “Research Highlights” feature on its website, covering timely and interesting economics research from its line of flagship scholarly publications.

Coverage is tailored to assist academics, policymakers, and journalists who may not have time to stay abreast of the latest published articles. AEA Research Highlights is also an accessible way for the general public to understand academic research via summaries, interactive links and graphs, photos, and presentations of lessons drawn from research.

AMERICAN ECONOMIC ASSOCIATION

See the latest AEA Research Highlights, or explore the archive at

www.aeaweb.org/research

Compliments of the American Economic Association

Webcasts of Selected Sessions from the 2020 AEA Annual Meeting

Now available on the AEA Website

AMERICAN
ECONOMIC
ASSOCIATION

January 3, 2020

Carbon Tax Policy

Presiding: Joseph E. Aldy, *Harvard University*

Measuring the Macroeconomic Impact of Carbon Taxes
Gilbert Metcalf, *Tufts University* and *NBER* and James Stock, *Harvard University*

The Fiscal Costs of Climate Change
Lint Barrage, *University of California-Santa Barbara* and *NBER*

The Welfare Implications of Carbon Price Certainty
Joseph E. Aldy, *Harvard University* and *NBER* and Sarah Armitage, *Harvard University*

Meredith Fowlie, *University of California-Berkeley*
Roberton Williams III, *University of Maryland* and *NBER*
Garth Heutel, *Georgia State University* and *NBER*

AEA/AFA Joint Luncheon

Nudges are Not Enough: The Case for Price-Based Paternalism
David Laibson, *Harvard University*, introduced by David Hirshleifer, *University of California-Irvine*

How Can Economics Solve Its Race Problem?

Presiding: Janet Yellen, *Brookings Institution*

Ebonya Washington, *Yale University*
Randall Akee, *University of California-Los Angeles*
Cecilia Conrad, *Pomona College*
Trevon Logan, *Ohio State University*
Edward Miguel, *University of California-Berkeley*
Marie T. Mora, *University of Missouri-St. Louis*

AEA Richard T. Ely Lecture: Gender in the 21st Century

Marianne Bertrand, *University of Chicago*, introduced by Janet Yellen, *Brookings Institution*

January 4, 2020

Alan Krueger's Contributions to Economics

Presiding: David Card, *University of California-Berkeley*

Alexandre Mas, *Princeton University*
Diane Whitmore Schanzenbach, *Northwestern University*
Lawrence Katz, *Harvard University*
Lisa Lynch, *Brandeis University*
David Card, *University of California-Berkeley*

Economics for Inclusive Prosperity (EfIP)

Presiding: Dani Rodrik, *Harvard University*

Can Economics Rise to the Challenge?
Suresh Naidu, *Columbia University*, Dani Rodrik, *Harvard University*, and Gabriel Zucman, *University of California-Berkeley*

Economic Foundations of Real Freedom for All
Samuel Bowles, *Santa Fe Institute* and Wendy Carlin, *University College London*

The Political Limits of Economics
Luigi Zingales, *University of Chicago*

Nobel Laureate Luncheon honoring the 2018 Nobel Laureates William Nordhaus (Yale University) and Paul Romer (New York University)

Presiding: Janet Yellen, *Brookings Institution*

Lint Barrage, *University of California-Santa Barbara*
Robert Gordon, *Northwestern University*
George Akerlof, *University of California-Berkeley*
Chad Jones, *Stanford University*

Deaths of Despair and the Future of Capitalism

Presiding: Angus Deaton, *Princeton University*

Anne Case, *Princeton University*
Robert D. Putnam, *Harvard University*
Raghuram Rajan, *University of Chicago*
Kenneth Rogoff, *Harvard University*

AEA Awards Ceremony

Presiding: Janet Yellen, *Brookings Institution*

AEA Presidential Address: Twenty-first Century Monetary Policy

Ben Bernanke, *Brookings Institution*, introduced by Janet Yellen, *Brookings Institution*

January 5, 2020

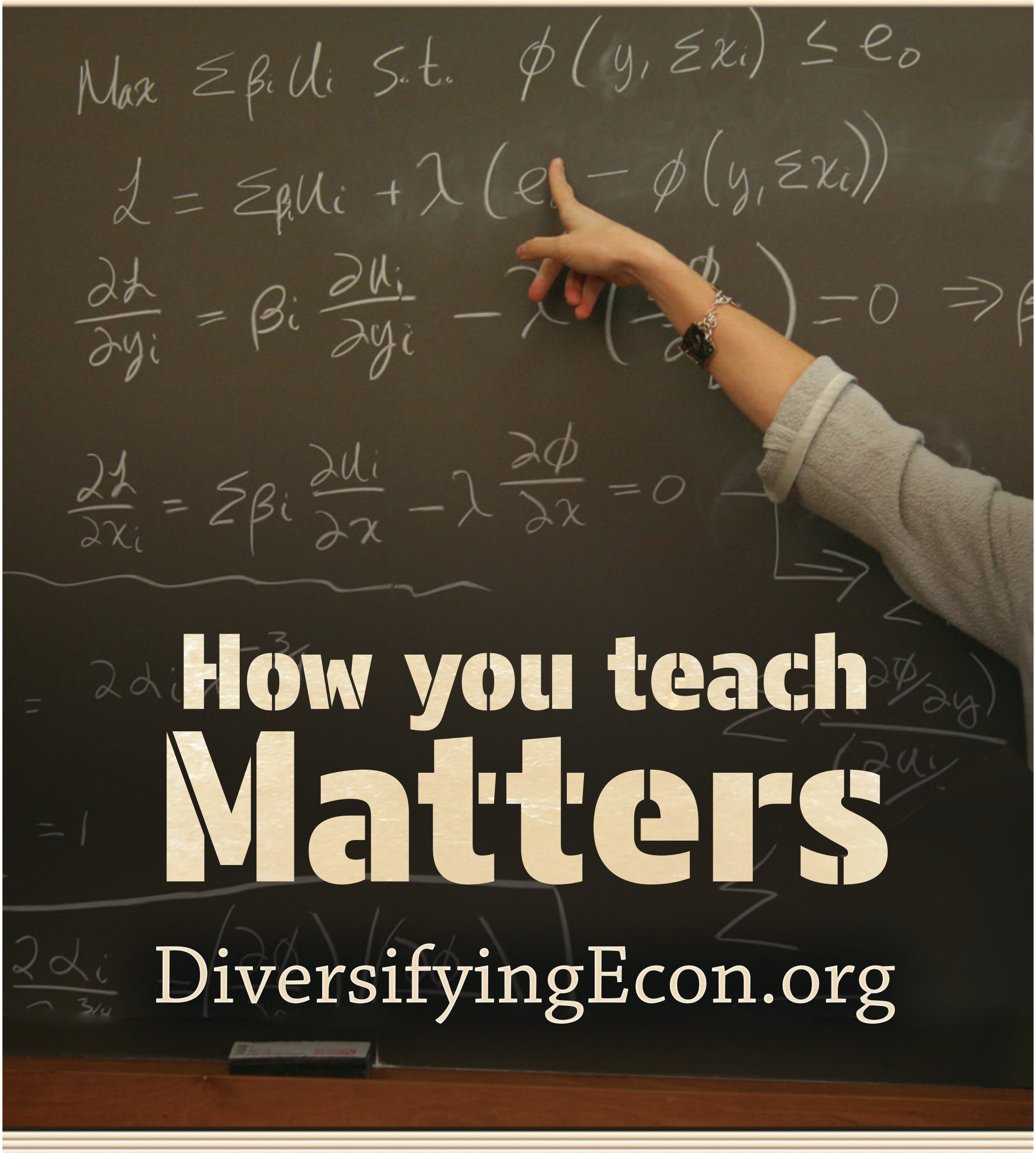
Japanification, Secular Stagnation, and Fiscal and Monetary Policy Challenges

Presiding: Adam Posen, *Peterson Institute for International Economics*

Mario Draghi, *European Central Bank*
Lawrence Summers, *Harvard University*
Janet Yellen, *Brookings Institution*
Adam Posen, *Peterson Institute for International Economics*

2020 AEA Continuing Education webcasts also available

www.aeaweb.org/conference/webcasts/2020


Max Σβ_iU_i s.t. φ(y, Σx_i) ≤ e_0
How you teach
Matters
DiversifyingEcon.org

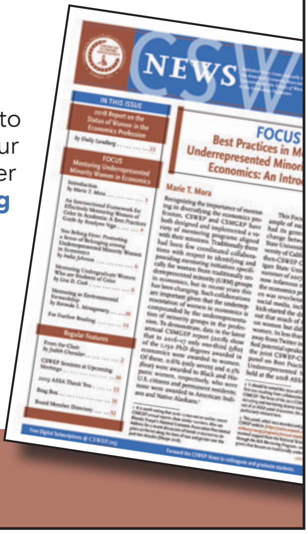
Committee on the Status of Women in the Economics Profession
AMERICAN ECONOMIC ASSOCIATION
Established 1971

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 representation of women in economics
- Offers mentoring workshops for junior faculty
- Conducts programs at the AEA and regional meetings
- Co-sponsors the summer economics fellows program
- Publishes a newsletter with professional development advice



Sign up to
receive our
newsletter
info@cswep.org

www.CSWEP.org

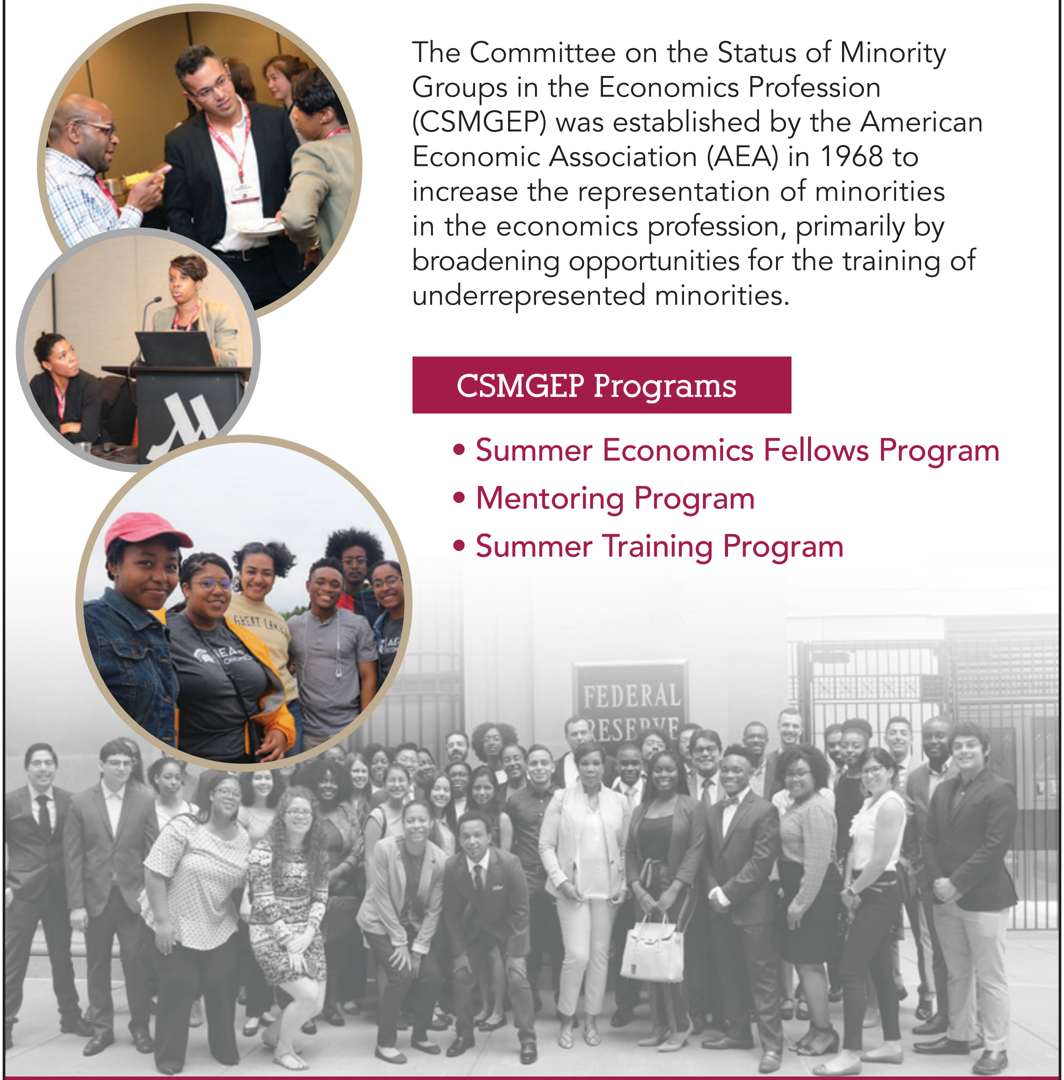
@AEACSWEP

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

www.csmgep.org

AEA RCT Registry
MOST RECENTLY REGISTERED TRIALS

Designing theories for the real world
The AEA interviews 2018 John Bates Clark Medalist Parag Pathak
American Economic Association

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.

FSC
www.fsc.org
MIX
Paper from responsible sources
FSC™ C132124

The Journal of

Economic Perspectives

Spring 2020, Volume 34, Number 2

Symposia

One Hundred Years of Women's Suffrage

Perspectives on Racial Discrimination

How Taxes Affect Location Choices

The Departure of Communism

Articles

AMERICAN
ECONOMIC
ASSOCIATION