The Journal of Economic Perspectives

A journal of the American Economic Association

Summer 2026

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.

Composed by American Economic Association Publications, Pittsburgh, Pennsylvania, USA.
Printed at Sheridan Press, Hanover, Pennsylvania, USA.

THE JOURNAL OF ECONOMIC PERSPECTIVES (ISSN 0895-3309), Summer 2026, Vol. 40, No. 3. The JEP is published quarterly (February, May, August, November) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203-2418. For details and further information on the AEA go to https://www.aeaweb.org/. Periodicals postage paid at Nashville, TN, and at additional mailing offices.

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

The Journal of
Economic Perspectives

Contents

Volume 40 • Number 3 • Summer 2026

Statement of Purpose

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

Policy on Data Availability

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

Policy on Disclosure

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

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 3–22

AI and Our Economic Future

Charles I. Jones

Artificial intelligence (AI) will likely be the most transformative technology of the modern era. Earlier technologies such as electricity, semiconductors, and the internet profoundly reshaped economic activity and increased living standards throughout the world. In some sense, AI is simply the latest in this line of general purpose technologies and at a minimum should continue the economic transformation that has been ongoing for the past century.

However, a case can be made that this time is different. Automating intelligence itself—that is, creating a technology that can perform any cognitive task that humans can do today—would have broad ramifications. Previous technological advances replaced humans in many physical tasks or in a narrow set of cognitive tasks, making human intelligence more valuable. But what if machines—AI for cognitive tasks and AI plus advanced robots for physical tasks—can perform every task a human can, at least as cheaply? While this outcome is by no means guaranteed in the next several decades, a number of observers expect that it will occur (Amodei 2024; Aschenbrenner 2024; Brynjolfsson, Korinek, and Agrawal 2025; Kokotajlo et al. 2025).

What does economics have to say about this possibility, and what might our economic future look like?

■ *Charles I. Jones is the Stanco25 Professor of Economics, Stanford Graduate School of Business, Stanford, California. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. After this paper was completed and in galley proofs, Jones was contacted by Anthropic to discuss the possibility of joining the company. Starting on June 30, 2026, he will be on leave from both Stanford and the NBER and working at Anthropic as a paid employee. His email address is chad.jones@stanford.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20261505.

We begin by outlining two scenarios for the impact of artificial intelligence on the economy: one in which AI drastically accelerates economic growth and another in which AI is "business as usual." Both scenarios are possible, and both scenarios have features that are informative about the future consequences of AI.

The future probably lies somewhere in between these two scenarios, but where? To shed insight on this question, we describe task-based models of economic growth in which tasks are complementary—a "weak link" framework. Just as a chain is only as strong as its weakest link, the economy may be limited by whatever tasks are not yet automated—that is, by tasks that are performed by slowly improving humans rather than by rapidly improving machines. This feature tends to slow the transformative effects of artificial intelligence. But "slow" does not mean "eliminate," and AI is likely to dramatically transform the economy over a time horizon measured in decades. In the United States, per capita GDP growth has averaged 2 percent per year for the past 150 years and been remarkably stable. Nevertheless, if AI eventually automates away nearly all the weak links in the economy, economic growth could accelerate significantly, with rates potentially exceeding 5 percent per year (Jones and Tonetti 2026).

Powerful and transformative AI also entails risks; it is a double-edged sword. What happens to labor markets, inequality, and the political equilibrium when machines can perform nearly every task that humans can? More speculatively, transformative AI enables potentially catastrophic risks, for example through cyberattacks, biological vectors, and autonomous weapons. As the leading AI labs have emphasized, these possibilities deserve serious consideration.

It is surely obvious but also important to acknowledge that no one knows what the consequences of automating intelligence will be. This essay provides one perspective based on economic research, but the perspective is inherently speculative and will likely prove to be incorrect on many dimensions. Nevertheless, I have found that the tools of economics are helpful in providing guideposts and insights as we consider how AI will impact our economic future.

Two Scenarios for the Future of Growth

To begin, let me sketch out two possible futures for the macroeconomy. These scenarios are intentionally extreme, at both ends of the spectrum. In the first, artificial intelligence has massive economic impacts and dramatically accelerates economic growth. In the second, AI is more "business as usual" or a "normal technology" to use the phrase coined by Narayanan and Kapoor (2025). Both scenarios are possible outcomes in my view; I will explain later how I think of where in between these two outcomes our future may lie.

Each scenario will describe changes that could occur over 25 or 30 years. Thus, the time period I have in mind is roughly one generation: what might the world look like as a child born today grows into an adult, or as a young adult now entering the workforce moves through their career?

AI Accelerates Economic Growth

What would the next several decades be like if artificial intelligence accelerates economic growth and has truly profound macroeconomic consequences? Scenarios along these lines have been presented by various authors.[1] While the scenarios are inherently speculative, I find two aspects to be intriguing. First, there is already evidence for their early stages, and second, each step does not seem impossible.

The scenarios typically begin with artificial intelligence raising the productivity of software engineers, which is a process that is already solidly under way. As one of a parade of similar examples, when Anthropic introduced Claude Opus 4.5 in November 2025, the firm highlighted Claude's exceptional performance on a two-hour take-home exam they give to prospective software engineering hires: the AI model scored higher than any human candidate ever (Anthropic 2025).

Epoch AI estimates that the amount of "effective compute" (that is, total computing power adjusted for the quality of algorithms and software) used to train artificial intelligence models is rising annually by a factor of ten: a factor of 4 from more and better computer chips and a factor of 2.5 from better algorithms (Ho et al. 2024; Epoch AI 2026). These investments result in rapid improvements: for example, METR (2026) has developed a test of the extent to which AI tools can autonomously carry out programming projects. In 2020, ChatGPT 3 could complete a task that would take a human nine seconds with a 50 percent probability of success. By June 2024, AI could solve an eleven-minute project with this accuracy. Two years later in mid-2026, the corresponding 50 percent success time had risen to twelve hours. The METR time horizon for software engineering has been doubling roughly every five to seven months. There is some suggestion that progress has accelerated, with the doubling time falling to just four months since the reasoning models (for example, o1-preview) were introduced in late 2024 (Emberson 2025).

One of the frontier uses of artificial intelligence models is creating AI agents, models adept at performing tasks currently done by humans on computers. Beyond coding, these include writing and editing documents, synthesizing reports, building spreadsheet models, developing slide presentations, making videos, brainstorming ideas, suggesting science experiments, and proving theorems in mathematics. In many fields and for many cognitive tasks, AI already performs at the level of a competent graduate student. Given the rapid progress over the past five years, AI models a decade from now may be able to accomplish nearly any of the tasks that talented humans can currently accomplish on computers.

One such task is doing artificial intelligence research. AI models may themselves, with no human involvement, eventually be able to create better AI models, a phenomenon known as *recursive self-improvement*. Combining this flywheel effect with rapid increases in computing power, it is possible that sometime in the next decade—and perhaps much sooner—we will have access to what Dario Amodei, CEO of Anthropic, has called "a country of geniuses in a data center" (Amodei 2024).

[1] For example, see Davidson (2021), Kokotajlo et al. (2025), Erdil et al. (2025), Cunningham (2025), and Davidson, Halperin, Houlden, and Korinek (2026).

Millions of virtual geniuses in a data center, each running much faster than the speed of the human mind, could accelerate the discovery of new ideas. We have already seen new discoveries by AI models: one vivid example is AlphaFold, the model that "solved" the problem of how to determine the three-dimensional structure of more than 200 million proteins given only the sequence of amino acids that defines them, leading to the 2024 Nobel Prize in chemistry for Demis Hassabis and John Jumper of Google DeepMind. An economy of AI geniuses could design new pharmaceuticals and predict which drugs are likely to do best in clinical trials with minimal side effects. They could synthesize the existing research literature and advise scientists on what lab experiments to run in fields ranging from biochemistry to materials science to nuclear fusion and clean energy.

In the longer run, genius-level artificial intelligence models could use virtual simulations and human-run experiments to design and build better robots, iterating until exceptional performance is achieved. Eventually, this flywheel could result in AI models that can perform nearly all cognitive tasks and robots that are designed, produced, and managed by AI that can perform nearly all physical tasks. If this scenario were to come to pass, even over the next half century, the gains in productivity and living standards would surely be tremendous. In standard growth models, this outcome would produce accelerating economic growth; in some cases, this acceleration could lead to growth rates exceeding 10 percent per year or more (Aghion, Jones, and Jones 2019; Jones and Tonetti 2026).

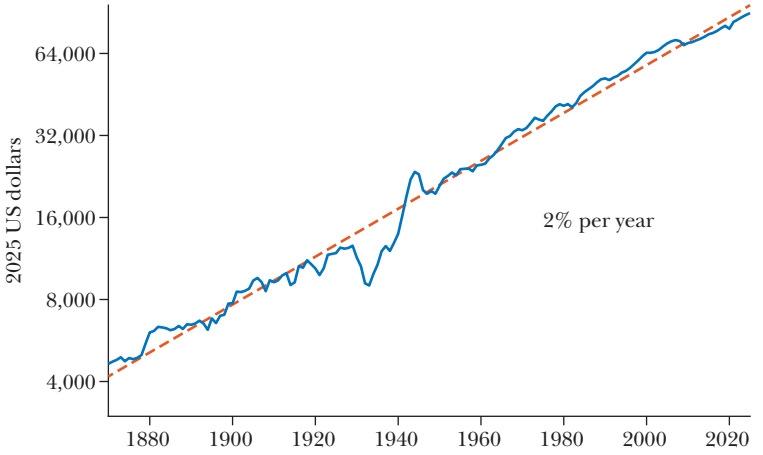
AI as "Business as Usual"

The US economy in the last half century has experienced dramatic increases in computing power, as illustrated by Moore's Law. Back in 1965, Gordon Moore, who was then director of research at Fairchild Semiconductor, wrote a paper predicting that the density of transistors on a computer chip would double every two years, a prediction that has largely been borne out across six decades (Moore 1965). The tremendous fall in the price of computing has brought technological miracles like the internet, but has not led to a sustained boost in the US growth rate. Is there also a scenario in which artificial intelligence leads to "business as usual" for future growth?

Figure 1 displays my favorite graph in economics: real GDP per person in the United States for the past 150 years. On a log scale, the time series is roughly a straight line. Indeed, Moore's Law and US GDP per capita growth are in some sense duals, illustrating the "rule of 70" that the product of the growth rate and doubling time under constant exponential growth equals 70: transistor density doubles every two years and grows at 35 percent per year while per capita US GDP grows at 2 percent per year and doubles every 35 years.

Consider the astounding innovations that underlie the graph. In the 1870s, Thomas Edison's experiments with electric lighting were just getting underway. Fifty years later, electrification had transformed the economy, both in factories and in city life. Throughout the 150 years, innovations such as the internal combustion engine, airplanes, vacuum tubes, antibiotics, transistors, semiconductors, personal computers, and the internet profoundly changed living standards. Many of these innovations are

Figure 1
Average US Income per Person



Source: Data are from US Bureau of Economic Analysis (2025) since 1929 and from Barro and Ursúa (2010) for 1870 to 1928.
Note: Average income per person in the United States rose at a remarkably steady rate of 2 percent per year for 150 years even while amazing technologies such as electricity and information technology transformed economic activity.

what economic historians call "general purpose technologies," whose transformative effects extend throughout the economy. Many also automated some of the tasks involved in creating new ideas—say, through improvements in scientific tools and equipment—raising the productivity of research and idea generation. Yet apparently none of these innovations changed the long-run growth rate of the US economy.

How can we understand this disconnect? One natural hypothesis is that within any technology field, ideas get harder to find (Bloom, Jones, Van Reenen, and Webb 2020). The steam engine runs out of steam. Without the discovery of the next general purpose technology, one might expect economic growth to slow down. From this perspective, each of these new general purpose technologies did indeed raise the growth rate of the economy in a counterfactual sense: but for their invention, the counterfactual is that growth would have slowed. The continued emergence of amazing new technology classes is what made sustained growth at 2 percent per year possible. Perhaps artificial intelligence is just the latest general purpose technology that lets 2 percent growth continue for several more decades. Notice that even in this scenario, AI potentially has large transformative effects, because the counterfactual is one in which growth would otherwise have slowed.

Another complementary lesson from economic history is that it can take decades for a new innovation to diffuse throughout the economy. David (1990) first made this point in the context of the steam engine and the electricity-generating dynamo. Robert Solow famously quipped in 1987: "You can see the computer age

everywhere but in the productivity statistics" (Solow 1987). Profound new technologies require many complementary innovations before they can take full effect. In the case of electricity, factories that had been driven by steam power needed to reorganize and even relocate production before they could realize the full productivity gains from small electric motors that could be placed throughout the factory. Organizational, product, and process changes needed to be implemented to take advantage of information technology (Brynjolfsson and Hitt 2000). For the new artificial intelligence, the lesson from economic history is that the effects on GDP and productivity may take decades to realize their full transformative effects.

Comparing the Scenarios

Both of the scenarios I have laid out have merit. The "accelerating growth" scenario is based on direct evidence about the rapidly evolving capabilities of artificial intelligence. Theoretical models of economic growth in which AI automates most of the tasks in the economy can formally produce explosive economic growth (Aghion, Jones, and Jones 2019). On the other hand, automation of production has been ongoing for more than 200 years, and transformative innovations such as electricity, semiconductors, and the internet have coexisted with remarkably stable economic growth at 2 percent per year.

In either scenario, the effects of artificial intelligence are large and profound, just as they were with electricity and the internet. Still, exactly how large and how profound is far from clear. Moreover, the future is likely to lie somewhere in between; substantial uncertainty surrounds the future effects of AI on the macroeconomy. We next turn to some theoretical insights that may shed light on these effects.

Weak Links and the Future of Economic Growth

Task-based models provide key insights to help us understand the effects of automation in general, and AI-style automation in particular.[2] In these models, production depends on the successful completion of a number of tasks. Initially, tasks are performed by labor. Automation occurs when firms figure out how to use machines in place of labor to perform particular tasks more cheaply. As Zeira (1998) emphasizes, automation has been going on at least since the Industrial Revolution of the nineteenth century, and actually for hundreds of years. A classic nineteenth-century example is the replacement of labor in weaving textiles by mechanical looms. In the twentieth century, cars and trucks replace people and horses in transportation, tractors replace people in many agricultural tasks, and electronic computers replace human computers in performing calculations.

In studying the effect of automation on economic growth, Aghion, Jones, and Jones (2019) focus on the case in which the production function depends on tasks

[2]Useful starting points in this literature include Autor, Levy, and Murnane (2003), Acemoglu and Restrepo (2018), Hémous and Olsen (2022), B. Jones and Liu (2024), and Korinek and Suh (2024).

that combine with an elasticity of substitution less than one. That is, the model is one of "weak links." Each task can be thought of as a link in the chain, and all tasks are essential to production. Similarly, just as a chain is only as strong as its weakest link, production is constrained by the weakest links—the least productive tasks (Kremer 1993; Jones 2011).

An iconic illustration of weak links arose with the accident that destroyed the space shuttle Challenger in 1986. The post-crash investigation found that the key design flaw was in the O-rings: doughnut-shaped pieces of rubber that are designed to create a seal between two interlocking parts. Challenger launched during record-low temperatures causing the O-rings to become brittle and leak, which caused the space shuttle to explode. Partly inspired by this episode, Kremer (1993) developed his O-ring model of economic development.[3]

The consequences of weak links need not be so dramatic, and potential weak links seem pervasive in production. Consider the creation and production of a new state-of-the-art smartphone. The smartphone must be designed, the chips must be manufactured to exact specifications, a myriad of sophisticated parts must be coordinated to arrive in the right place at the right time, the smartphone needs to be delivered to the retail store, advertising, marketing, and customer management must be arranged, and so on. When any of these tasks are hindered, the value of the enterprise gets reduced considerably.

Jones and Tonetti (2026) use the weak link framework to explain how the presence of weak links can limit the growth that emerges from artificial intelligence. A simple example illustrates how this works. Suppose that final output is produced by combining two tasks: $Y = F(X_{easy}, X_{hard})$. The "easy" task is easily automated, while the "hard" task is hard to automate. To keep things simple, let's suppose that $X_{hard} < X_{easy}$—the hard task is also harder to produce and so constitutes the smaller input.

As in much of the task literature, assume this production function features a constant elasticity of substitution (CES). It simplifies things considerably to further assume that the elasticity of substitution is equal to 1/2. In this case, output is proportional to the harmonic mean:[4]

$$\frac{1}{Y} = \frac{1}{X_{easy}} + \frac{1}{X_{hard}}.$$

[3]Kremer's O-ring model is based on a Leontief production function, that is, a CES production function with a zero elasticity of substitution. The weak link production function allows the importance of weak links to be a parameter, as discussed in the next several paragraphs and the next footnote.

[4]The standard CES production function in this case would be

$$Y = \left(X_{easy}^{\frac{\sigma-1}{\sigma}} + X_{hard}^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}.$$

Setting $\sigma = 1/2$ delivers the expression in the text. The standard arithmetic mean occurs if $\sigma = \infty$ so the exponents become one (we are ignoring the 1/2 multiplier that is needed to turn the sum into the mean). The geometric mean (Cobb-Douglas) occurs when $\sigma = 1$, although this is hard to see because the exponents become 0 and ∞. The harmonic mean is just another way of measuring the central tendency; it puts more weight on the smaller values—that is, on the weak links.

In this formulation, the hard task is the weakest link and it limits total production. That is, if $X_{easy} = \infty$, then $Y = X_{hard}$, so with less than infinite X_{easy}, we have that $Y < X_{hard}$. Notice that this also means that even with infinite amounts of some input, overall production remains finite—again because we are limited by the weakest link. These results generalize to a production function involving many tasks. In this sense, it becomes clearer how automating many tasks might not lead to huge output gains: output is always constrained by the weakest links that are not yet automated.

To see a familiar example of this phenomenon, just look at your computer. You have on your desk roughly 100 million times the computing power that was available on the best computers from the early 1970s. But you and I are not 100 million times more productive than the economists of that time. Matrices get inverted much more rapidly, but we still have to decide what data to put in the matrix, what the structure of the economic model is, and what theory we are testing. In other words, our production is limited by the weak links.

Now return to the harmonic mean example and suppose that we choose units so that initial output is $Y_{initial} = 1$. What happens if we then fully automate the "easy" task? In fact, the cleanest result comes from supposing we become so good at the easy task that $X_{easy} = \infty$; it is no longer a bottleneck in any way. In that case, it is straightforward to show that output with infinite amounts of X_{easy} is given by a simple formula: $1/(1 - s)$ where s is the original spending share on the easy task, assuming competitive markets.[5] That is, infinitely automating a task that costs the economy a fraction of GDP given by s raises output by the factor $1/(1 - s)$.

Next, recall the "artificial intelligence accelerates economic growth" scenario. One of the first pillars of that scenario is that AI automates software production. How large an effect would this have on the economy? Well, even with *infinite* amounts of software doing the existing set of tasks that software performs today, output would rise by $1/(1 - s)$. Spending on software accounts for around 2 percent of GDP, and for small s, $1/(1 - s) \approx 1 + s$. In other words, having access to infinite amounts of the tasks that software performs today would only raise GDP by around 2 percent. Why is the gain so small? Weak links.

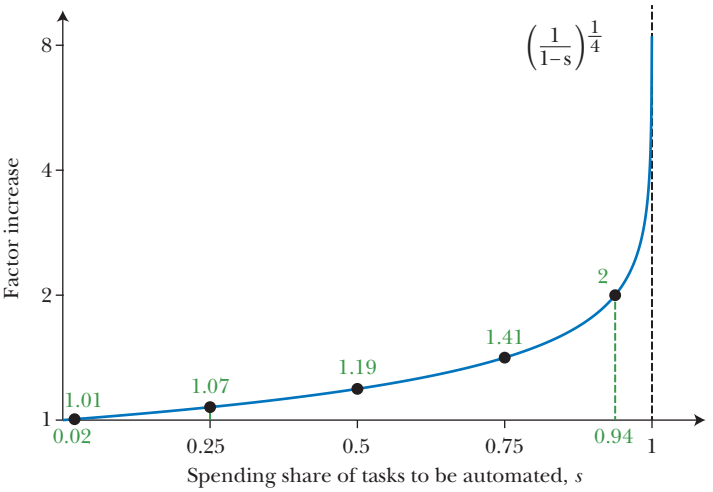
In the more general case where the elasticity of substitution—the weak link parameter—is σ, output after infinitely automating the share s of GDP is given by

$$\left(\frac{1}{1-s}\right)^{\frac{\sigma}{1-\sigma}}.$$

When $\sigma = 1/2$, this reduces to the simple case where the exponent is one. To the extent that σ is smaller, weak links are more important. For example, if $\sigma = 0.2$, then the exponent becomes $0.2/0.8 = 1/4$ and the output gains from infinite automation will be even smaller.

[5] Jones and Tonetti (2026) derive this formula for a more general constant-elasticity-of-substitution setup; see also B. Jones (2025).

Figure 2
GDP per Person with Infinite Automation



Source: Jones and Tonetti (2026).
Note: The figure shows the factor increase in GDP per person from automating with infinite productivity tasks that have an initial cost share of GDP equal to s, assuming an elasticity of substitution equal to 0.2.

Figure 2 illustrates this formula graphically. That is, the figure shows the proportional gain from infinitely automating more and more tasks for the case in which weak links are significant: $\sigma = 0.2$, which is the value chosen in Jones and Tonetti (2026) based on a range of evidence. The horizontal axis is s, the share of GDP that gets automated. There are two key messages from the figure. First, as the share of the economy that gets automated increases, the gains are larger. Indeed, if 100 percent of the economy is automated, GDP would be infinite. This is an intuitive explanation for how the gains from automation can ultimately be very large—as in the "AI accelerates growth" scenario considered earlier.

The second message from the figure is that weak links can significantly slow the gains from automation. In the "artificial intelligence accelerates growth" scenario, the next step after automating software is that AI automates all cognitive work that can be performed on a computer. To see the consequences of this, consider the case in which $s = 1/2$ so that 50 percent of GDP gets infinitely automated. According to Figure 2, this would raise GDP by just 19 percent. The reason is once again weak links. With $\sigma = 0.2$, having infinite amounts of half the tasks in the economy has relatively small effects because we are limited by the remaining tasks that are performed by humans. We are still using people to educate our children, provide healthcare, run the government, manufacture goods, and provide hotels, restaurants, and travel experiences. Indeed, even just doubling income per person requires the infinite automation of 94 percent of GDP.

Of course, the set of tasks that can be automated in the economy is not frozen in time but instead evolves as we discover new ideas. Jones and Tonetti (2026) build a dynamic model in which the production of both goods and ideas gets progressively automated over time. More automation leads to more new ideas, which in turn lead to further automation, and new ideas are continually making machines more and more productive at the tasks they perform. On the other hand, at any point in time this flywheel dynamic is constrained by weak links.

The equation graphed in Figure 2 can be viewed as an extremely simplified version of that much richer model. In particular, suppose that we begin the economy at $s = 0$ so there is no infinite automation. And suppose that each year s increases by .01, so that after 100 years all tasks will be infinitely automated. In that case, the curve in Figure 2 can be interpreted as the time series of income per person over the next century. And because the graph is plotted on a log scale, the slope of the curve reflects the growth rate over time. The slope is relatively stable for the first 50 or 60 years—suggesting that growth does not accelerate by much. We might expect growth rates to be similar to the 2 percent per year growth shown in Figure 1. After 75 years, however, the slope starts to increase, and over the last 25 years the slope rises dramatically, and GDP becomes infinite after a century has passed.

What is the lesson? The positive feedback loop between automation and the discovery of new ideas does indeed accelerate growth, as in the first scenario considered earlier in this essay. However, the presence of weak links means that the acceleration is initially slow, as in the "business as usual" scenario. It is only after the vast majority of weak links have been automated away that the growth acceleration truly becomes evident. These same results emerge in the richer and fully dynamic model in Jones and Tonetti (2026). On the one hand, weak links slow the economic gains from artificial intelligence and automation; the consequences could easily be modest for the next decade or two. On the other hand, if AI is ultimately able to automate nearly every task that humans perform, even the model that emphasizes substantial weak links features a huge acceleration in economic growth.

An example that I find helpful in illustrating this point is self-driving cars. The first DARPA Grand Challenge was in 2004. Fifteen self-driving vehicles competed and none finished the 142-mile course; the furthest any vehicle traveled was 7.5 miles (DARPA 2014). Twenty-two years later in 2026, Waymo self-driving cars are easy to spot throughout San Francisco. The vehicles have traveled a total of 170.7 million miles without a human driver, and the record suggests they are much safer than most human drivers (Waymo 2026). And yet San Francisco is the exception, not the rule. In the vast majority of other places in the United States, you still cannot hail a self-driving car. This is another illustration of weak links (weather conditions, the long tail of edge cases, cost of sensor suites and computing, and so on) and the time horizon that is required for those links to be overcome. I suspect the same thing will be true about many other aspects of artificial intelligence interacting with the physical world. We will need multiple "nines" of reliability before we have robots managing our kindergarten classrooms.

A range of possibilities exist between the "artificial intelligence accelerates economic growth" and "AI is business as usual" scenarios, and other research has also considered these possibilities. Nordhaus (2021) asks whether an economic "singularity" of extraordinarily rapid growth is near and concludes that it is not. Acemoglu (2025) suggests that the macroeconomic impacts of AI may be quite modest over the next decade, raising growth in total factor productivity by less than 0.1 percentage points per year. Aghion and Bunel (2024) question some of the empirical choices made by Acemoglu and calculate a larger gain, with AI perhaps raising growth of total factor productivity by 0.7 percentage points per year. Jones and Tonetti (2026) suggest that the "business as usual" scenario may be most relevant for the next decade or two but that it is likely that AI does indeed ultimately accelerate economic growth in the long run.

Labor Markets, Inequality, and Meaningful Work

Many of the concerns about artificial intelligence arise from its potential effects on labor markets. This important topic is further from my expertise, so I will offer only a few scattered remarks. Others such as Acemoglu and Autor (2011), Acemoglu and Restrepo (2022) and Autor and Thompson (2025) have studied this question closely.[6]

A key insight comes from appreciating that jobs are bundles of tasks. Artificial intelligence can therefore reduce wages in some jobs but raise wages in others, leading to nuanced and unexpected results. A classic example comes from a remark in 2016 by Geoffrey Hinton—who would later share the 2024 Nobel Prize in Physics "for foundational discoveries and inventions that enable machine learning with artificial neural networks"—that medical schools should stop training radiologists because AI would soon be much better at reading medical imaging scans, putting radiologists out of work. Nearly ten years later, we have *more* radiologists rather than fewer, and salaries for radiologists have risen rather than fallen (Mousa 2025). It turns out that radiologists do more than just read scans, and AI tools complement those other skills by automating a fraction of the tasks that radiologists perform. In a weak-link world, automation will raise the task prices of the remaining weak links, which in turn can raise wages, at least as long as those weak links are themselves not automated away. Automation of *some* of the essential tasks performed by radiologists raises the productivity of radiologists overall. Of course, some jobs—perhaps even radiologists some day—may consist of bundles of tasks that are all automated, and the wages of those jobs will fall to the marginal cost of performing those tasks with machines and software, which may indeed be low.

[6] See also Althoff and Reichardt (2026) and Freund and Mann (2026).

One of the questions at the start of this essay was "What happens if AI plus robots can perform every task a human can, at least as cheaply?" A first observation is that the weak links framework described in the previous section suggests that this may take decades to occur rather than happening in the next few years. So there is perhaps time for the labor market to adjust, as the changes come gradually rather than suddenly.

Beyond the rate of adjustment, the total computing power of the economy at any point in time is finite, so the principle of comparative advantage comes into play. There will always be tasks that are scarce and that humans can perform; the question is simply at what price. But as the economy gets more productive, the wage for some tasks must rise (Caselli and Manning 2019; Restrepo 2025). Humans are not getting *worse* at performing tasks, and some tasks will remain scarce. None of this is to say that some workers cannot see their wages fall. Perhaps humans leave highly-paid occupations like software engineers and lawyers in order to take lower paying jobs that retain value precisely because they are performed by humans (for example, artists and musicians). The wage can fall when people switch jobs even when the wage for artists is always rising. Similarly, even if in some scenarios the share of GDP paid to labor is falling to zero, this does not mean that the wage is falling. For example, if total labor income rises at 3 percent per year while total capital income rises at 5 percent per year, the labor share will fall to zero. But labor incomes are rising. All of this is simply to reinforce that the labor market consequences of AI can be nuanced.

Another point worth appreciating is that a world in which artificial intelligence combined with robotics can automate nearly all tasks is a world of abundance. The "size of the pie" could become very large, as we have already discussed, and in principle the large increases in GDP could make everyone better off: it becomes a question of distribution. Historically, the main asset that many people have is their labor endowment, and renting this asset to an employer was (and is) the way that many people earned income. Thus, an important question is what economically valuable endowment people possess that will enable them to share in any riches that AI creates. Modern economies engage in substantial redistribution through a system of taxes and benefits, and large gains in GDP would make this kind of redistribution more affordable. Of course, the political economy of redistribution could easily become more complicated to the extent that wealth gets even more concentrated—say, if the financial benefits of discovering advanced AI flow to a small set of people. Questions around the distribution of income are likely to become very important in the future and deserve careful consideration. Trammell and Patel (2025) offer a readable starting point for discussing the potential broader consequences of AI for wealth inequality.

In most economic models, work is a "bad" rather than a "good"—which is why people must be paid to do it. However, for many people, and perhaps even for an increasing number over time, jobs provide one foundation for finding meaning in their lives. When AI is better than me at understanding the sources of economic growth and developing new growth models, where will I find meaning in my life?

In early 2026, ChatGPT 5.2 Pro was already better than me at solving many of the growth problems that I work on.[7]

I do not know the answer to this question, but metaphors I have found helpful in thinking about it are *retirement* and *summer camp*. Retirees with ample incomes and plentiful time seem to find meaning in life by spending time with friends, travel, new experiences, and so on. Perhaps in the future, my growth economist friends and I will gather in interesting locales and have artificial intelligence teach us the latest growth models that it has discovered. Most of what we enjoy about research is learning new things, and this will only get easier as AI improves. In the end, advanced AI will perhaps understand humanity better than we understand ourselves, and we can ask it for advice about how best to live a meaningful life in a brave new world.

Catastrophic Risk

Modern-day Prometheans like Sam Altman of OpenAI, Dario Amodei of Anthropic, Demis Hassabis of Google DeepMind, and Geoff Hinton, who is sometimes known as the "Godfather of AI" for his work on neural networks, were all early advocates of the incredible promise of artificial intelligence. But they share another important characteristic as well. Before 2020, they all warned about the significant potential risks of AI technologies: a common message was that AI may be a larger economic driver than electricity or the internet, but could also be more dangerous than nuclear weapons. Indeed, OpenAI was founded explicitly as a nonprofit so that it could focus on developing AI safely, avoiding pressures from market incentives to develop powerful AI models before safety measures were fully in place.[8]

Risks from artificial intelligence broadly fit into two categories. The first is associated with *bad actors*. Consider AI models in the next five to ten years. Extrapolating from recent gains, we could have amazing AI models that could master biochemistry and virology. A bad actor could ask such a model to develop the recipe for a novel virus that is more deadly and more contagious than Ebola, but that takes four weeks for symptoms to emerge. The modern world has so far managed to avoid the worst outcomes associated with nuclear weapons in part because only a small number of parties had access to the "red button." Advanced AI could mean we have millions of people with access to red buttons.

The second category of catastrophic risk is more speculative and might be called *alien intelligence*. We are growing new forms of intelligence that we do not understand. Suppose we found out tomorrow that an alien spacecraft was passing

[7]For example, see this problem of characterizing explosive economic growth: https://web.stanford.edu/people/chadj/explosivegrowthChatGPT5.2.pdf.

[8]For discussions of catastrophic risks from artificial intelligence and other sources, useful starting points are Bostrom (2002), Rees (2003), Posner (2004), Yudkowsky (2008), and Nielsen (2024). And of course OpenAI is no longer a pure nonprofit. Ord (2020), in Table 6.1, attempts to quantify other sources of existential risk as well. Unaligned AI is by far the largest risk, followed by engineered pandemics.

Pluto on its way to Earth. How would we feel? It would surely be exciting, but we might also feel some trepidation: when more advanced species or societies encounter less advanced ones historically, it does not usually end well for the less advanced party. Stuart Russell, a Berkeley computer science professor who is coauthor of one of the leading graduate textbooks on artificial intelligence, has offered a thought-provoking comment: "How do we retain power over entities more powerful than us, forever?"[9]

These risks are inherently speculative. For present purposes, a natural question is this: To what extent can economic analysis shed light on these risks given the inherent uncertainties? The remainder of this section discusses some progress economists have made along these lines.

The Oppenheimer Question

In anticipation of the first test of the atomic bomb, the scientists of the Manhattan Project faced potentially catastrophic risks: for example, what if the nuclear chain reaction continued unabated, igniting the atmosphere and potentially killing all life on Earth? The scientists considered this question, as explained in a now-declassified report by Konopinski, Marvin, and Teller (1946). They estimated the probability to be extremely low, and the Trinity test went forward. But how large would the risk have to be to avert the test?

Jones (2024) considers an analogous question with respect to artificial intelligence. Suppose on one side that AI has incredible benefits, raising economic growth to 10 percent per year and thus doubling the average standard of living every seven years. However, it comes with a one-time risk of killing everyone on the planet. How much risk would standard economic agents be willing to take in this case?

Several surprising results emerge. First, consider the situation in which agents have log utility and begin with living standards comparable to the US average. With log preferences, the marginal utility of consumption is proportional to the inverse of consumption—a doubling of consumption leads marginal utility to fall by 50 percent. In this case, it turns out that the agent would be willing to take any risk lower than 1-in-3 of killing everyone in order to get 10 percent growth. (I learned from this calculation that I do not have log utility!)

Next, consider a situation in which marginal utility falls with the *square* of consumption, which is a case often considered in macroeconomics. Now, the marginal utility of consumption falls twice as fast. The gain from 10 percent annual consumption growth is worth much less, and the acceptable risk plummets: agents are only willing to accept a 2.5 percent existential risk in exchange for growth. At US living standards, utility is relatively high already and the marginal utility of additional consumption is relatively low. It is not worth taking large risks of losing our high living standards to get gains in consumption that are not very valuable.

[9] From a panel presentation in Anton Korinek's CEPR AI online seminar, February 25, 2025.

The final surprise in Jones (2024) comes from recognizing that if artificial intelligence is remarkable enough to deliver 10 percent economic growth, it will probably also produce tremendous advances in medicine—perhaps curing cancer and heart disease, and offering a dramatic extension of healthy lifespans. Suppose that the new AI future cuts mortality rates in half (admittedly a large change, but then, so is the shift to 10 percent growth). The agents must now include in their calculations that in the good scenario, AI will not only improve consumption but will also dramatically reduce the chance of death and extend life expectancy. In this case, it turns out that even when the marginal utility of consumption falls rapidly, agents are willing to take large existential risks. The relevant tradeoff is not consumption versus death but rather death from cancer versus death from misaligned AI (in which case the marginal utility of consumption is no longer particularly relevant). Although agents care about not dying, they do not care about the cause of death. In this case, no matter how fast the marginal utility of consumption declines, standard economic agents are willing to take a 1-in-4 chance of killing everyone in order to cut mortality rates in half. Clearly, the health and medical benefits of artificial intelligence may be particularly valuable.

These thought experiments obviously ignore many important considerations. Perhaps we assign a value to humanity continuing over and above the selfish benefits considered in these examples. Nevertheless, the calculations may be informative, particularly about the types of artificial intelligence models that are most valuable—for example, medical innovations may be especially important.

How Much Spending to Mitigate Existential Risk?

Of course, just because people are *willing* to accept a certain amount of risk does not mean that we should do nothing. What if we can undertake actions to reduce the existential risk itself? How much should we be spending to mitigate existential risk (Jones 2025)?

A point emphasized by Ord (2020) and MacAskill (2022) is that an existential risk that ends humanity would arguably prevent the existence of thousands of future generations, constituting trillions of future people. When a generation stands at the precipice of a decision that will have substantial consequences for future generations, it arguably has a moral obligation to undertake large investments to limit the existential risk. This point is clearly important, but it also involves a moral argument that we should give substantial weight to people in distant future generations who do not yet exist—and indeed, may never exist.

Our recent experience with the COVID-19 pandemic suggests a very different motivation for spending to mitigate existential risk. In particular, in 2020, we each faced an imminent mortality risk of something like 0.3 percent from COVID-19. As a society, we responded by shutting down the economy and remaining in our homes, "spending" the equivalent of around 4 percent of GDP in the United States to mitigate this risk to current generations (Goolsbee and Syverson 2021; Fernández-Villaverde and Jones 2020). If one believes the catastrophic risks from artificial intelligence are at least this large, then by revealed preference perhaps we

should be spending an equivalent amount, even from a purely selfish standpoint that places no value on future generations.

A counterargument is that what we did in 2020 may not have been optimal. However, a simple calculation along the lines of Hall, Jones, and Klenow (2020) suggests that this consideration strengthens rather than weakens the argument. US government agencies implementing safety policies routinely use numbers on the order of $10 million or more for the value of life for an average American today (US Environmental Protection Agency 2026; US Department of Transportation 2026). To avoid a mortality risk of 1 percent, this value implies a willingness to pay of 1 percent $\times$ $10 million = $100,000. Average US GDP per person is around $90,000, so if projected over the entire population, this willingness to pay to reduce mortality risk by 1 percent is more than 100 percent of GDP. If the existential risk is realized in the next 10 to 20 years, an annual investment of 5–10 percent of income could be appropriate, at least if it would completely eliminate the risk.

This willingness to pay needs to be adjusted by a measure of the effectiveness of the mitigation spending, which for artificial intelligence has considerable uncertainty. However, some forms of mitigation could be effective. For example, many of the early efforts by DeepMind were *narrow* AI models such as AlphaFold. We could focus government support of AI research on narrow models that accelerate scientific research, especially in medicine, on the grounds that narrow AI may pose smaller risks. Alternatively, we could put greater emphasis on AI safety research. Algorithms that improve AI safety can be thought of as global public goods; there are surely large spillovers both across countries and over time from these types of ideas. Jones (2025) conducts various robustness checks and shows that even with a wide range of the effectiveness of mitigation spending, we are likely underinvesting in AI safety by a factor of 30 or more. Investments on the order of $100 billion per year—a third of a percent of US GDP—easily pass cost-benefit analysis criteria.

At current US levels of consumption, life is incredibly valuable and the marginal utility of consumption is correspondingly low. From a purely selfish standpoint—placing no weight on future generations—it is worth spending perhaps surprisingly large amounts of money to mitigate risks to life based on valuations that the US government uses every day.

An AI Arms Race and Policy

It feels as if there is a prisoners' dilemma element to artificial intelligence race dynamics at the moment. On the one hand, many of the leaders of the AI labs have historically warned about the potential risks associated with AI. On the other hand, these same leaders seem to be racing ahead to build data centers and advance the technology before safety problems are solved. One can imagine each lab individually saying: Everyone else is racing. If I slow down, that does not meaningfully change the existential risks. But if I race, then (1) maybe I will be safer than the others, and (2) enormous gains await the winner of the race if the catastrophic risks are not realized. The equilibrium is that everyone races, even though everyone may be better off if all slowed.

If the main parties understand the negative aspects to arms race dynamics, it raises the possibility that international cooperation, mediated and checked by third parties, could be possible. For example, the 1970 Treaty on the Non-Proliferation of Nuclear Weapons seems to have slowed the spread of nuclear weapons in the last half-century, and the 1987 Montreal Protocol on Substances that Deplete the Ozone Layer coordinated a global reduction in emissions of chlorofluorocarbons and other substances that diminished levels of ozone in the atmosphere.

Concluding Thoughts

How much will artificial intelligence change the world in the coming decades? A lesson from economic history is that general purpose technologies like the internet take multiple decades to have their full impact, and surely the same will be true of AI. Just because the effects have been modest for the past decade does not mean that the cumulative effect over half a century will be small.

The weak link framework helps us to understand this statement and to reconcile the two scenarios at the beginning of this essay. There is indeed a positive feedback loop between innovation and automation. Advances in artificial intelligence help us create new algorithms and new ideas, and these new ideas in turn help us advance AI and automation. This positive feedback loop provides an impetus for economic growth to accelerate. But weak links constrain the acceleration. We are not 100 million times more productive simply because we use today's computer chips instead of chips from the 1970s. Many other necessary tasks for production have not seen this kind of stunning improvement and therefore serve as weak links that constrain the economy's performance. But if AI and AI-run robots manage to automate nearly all of the economy, the positive feedback loop could indeed accelerate economic growth to unprecedented rates, possibly 5 percent per year or even higher.

The weak link view suggests that the substantial benefits from artificial intelligence might be delayed by decades as a large fraction of the weak links must be automated away. But the weak link view has the opposite implication for catastrophic risks. When a chain is only as strong as its weakest link, cutting one of the links is sufficient to do tremendous damage to the value of the chain. AI automation of software engineering in the near future may not lead to large benefits in the short term because of the other weak links. But powerful AI could do enormous damage to the economy by hacking the electricity grid or the financial system or helping a bad actor to engineer a deadly pandemic. Because of weak links, the benefits from AI may arrive gradually but some risks may arrive quickly. It would be prudent to spend the intervening time making preparations for the potentially large consequences for labor markets, inequality, and catastrophic risk.

■ *The author is grateful for helpful comments from David Autor, Basil Halperin, Ben Jones, Tom Houlden, Pete Klenow, Jeffrey Kling, Narayana Kocherlakota, Jonathan Parker, Timothy Taylor, Chris Tonetti, Phil Trammell, and Heidi Williams.*

References

Acemoglu, Daron. 2025. "The Simple Macroeconomics of AI." *Economic Policy* 40 (121): 13–58.

Acemoglu, Daron, and David Autor. 2011. "Skills, Tasks and Technologies: Implications for Employment and Earnings." In Handbook of Labor Economics, Vol. 4, edited by David Card and Orley Ashenfelter, 1043–171. North-Holland.

Acemoglu, Daron, and Pascual Restrepo. 2018. "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment." *American Economic Review* 108 (6): 1488–542.

Acemoglu, Daron, and Pascual Restrepo. 2022. "Tasks, Automation, and the Rise in U.S. Wage Inequality." *Econometrica* 90 (5): 1973–2016.

Aghion, Philippe, and Simon Bunel. 2024. "AI and Growth: Where Do We Stand?" Unpublished.

Aghion, Philippe, Benjamin F. Jones, and Charles I. Jones. 2019. "Artificial Intelligence and Economic Growth." In *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 237–82. University of Chicago Press.

Althoff, Lukas, and Hugo Reichardt. 2026. "Task-Specific Technical Change and Comparative Advantage." CESifo Working Paper 12403.

Amodei, Dario. 2024. "Machines of Loving Grace: How AI Could Transform the World for the Better." Unpublished.

Anthropic. 2025. "Introducing Claude Opus 4.5." November 2025. https://www.anthropic.com/news/claude-opus-4-5.

Aschenbrenner, Leopold. 2024. "Situational Awareness: The Decade Ahead." June 2024. https://situational-awareness.ai/.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." *Quarterly Journal of Economics* 118 (4): 1279–333.

Autor, David, and Neil Thompson. 2025. "Expertise." *Journal of the European Economic Association* 23 (4): 1203–71.

Barro, Robert J. and José F. Ursúa. 2010. *Barro-Ursúa Macroeconomic Data*. https://barro.scholars.harvard.edu/data_sets.

Bloom, Nicholas, Charles I. Jones, John Van Reenen, and Michael Webb. 2020. "Are Ideas Getting Harder to Find?" *American Economic Review* 110 (4): 1104–44.

Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9.

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

Brynjolfsson, Erik, Anton Korinek, and Ajay K. Agrawal. 2025. "A Research Agenda for the Economics of Transformative AI." NBER Working Paper 34256.

Caselli, Francesco, and Alan Manning. 2019. "Robot Arithmetic: New Technology and Wages." *American Economic Review: Insights* 1 (1): 1–12.

Cunningham, Tom. 2025. "Economics and Transformative AI." October 2. https://tecunningham.github.io/posts/2025-09-19-transformative-AI-notes.html.

DARPA. 2014. "The DARPA Grand Challenge: Ten Years Later." Defense Advanced Research Projects Agency, Tactical Technology Office, March 13. https://www.darpa.mil/news/2014/grand-challenge-ten-years-later.

David, Paul A. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *American Economic Review* 80 (2): 355–61.

Davidson, Tom. 2021. "Could Advanced AI Drive Explosive Economic Growth?" Coefficient Giving, June 25. https://coefficientgiving.org/research/could-advanced-ai-drive-explosive-economic-growth/.

Davidson, Tom, Basil Halperin, Thomas Houlden, and Anton Korinek. 2026. "Is Automating AI Research Enough for a Growth Explosion?" Unpublished.

Emberson, Luke. 2025. "Frontier AI Capabilities Accelerated in 2024." Epoch AI Substack, December 23. https://epochai.substack.com/p/frontier-ai-capabilities-accelerated.

Epoch AI. 2026. *Data on AI Models*. https://epoch.ai/data/ai-models?view=graph&tab=notable (accessed December 18, 2025).

Erdil, Ege, Andrei Potlogea, Tamay Besiroglu, Edu Roldan, Anson Ho, Jaime Sevilla, Matthew Barnett, Matej Vrzala, and Robert Sandler. 2025. "GATE: An Integrated Assessment Model for AI

Automation". https://arxiv.org/abs/2503.04941.
Fernández-Villaverde, Jesús, and Charles I. Jones. 2020. "Macroeconomic Outcomes and COVID-19: A Progress Report." *Brookings Papers on Economic Activity* 51 (2): 111–66.
Freund, Lukas B., and Lukas F. Mann. 2026. "Job Transformation, Specialization, and the Labor Market Effects of AI." CESifo Working Paper 12072.
Goolsbee, Austan, and Chad Syverson. 2021. "Fear, Lockdown, and Diversion: Comparing Drivers of Pandemic Economic Decline 2020." *Journal of Public Economics* 193: 104311.
Hall, Robert E., Charles I. Jones, and Peter J. Klenow. 2020. "Trading Off Consumption and COVID-19 Deaths." *Federal Reserve Bank of Minneapolis Quarterly Review* 42 (1): 2–13.
Hémous, David, and Morten Olsen. 2022. "The Rise of the Machines: Automation, Horizontal Innovation, and Income Inequality." *American Economic Journal: Macroeconomics* 14 (1): 179–223.
Ho, Anson, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. 2024. "Algorithmic Progress in Language Models." Preprint, arXiv. https://doi.org/10.48550/arXiv.2403.05812.
Jones, Benjamin. 2025. "Artificial Intelligence in Research and Development." NBER Working Paper 34312.
Jones, Benjamin F., and Xiaojie Liu. 2024. "A Framework for Economic Growth with Capital-Embodied Technical Change." *American Economic Review* 114 (5): 1448–87.
Jones, Charles I. 2011. "Intermediate Goods and Weak Links in the Theory of Economic Development." *American Economic Journal: Macroeconomics* 3 (2): 1–28.
Jones, Charles I. 2024. "The AI Dilemma: Growth versus Existential Risk." *American Economic Review: Insights* 6 (4): 575–90.
Jones, Charles I. 2025. "How Much Should We Spend to Reduce A.I.'s Existential Risk?" NBER Working Paper 33602.
Jones, Charles I., and Christopher Tonetti. 2026. "Past Automation and Future A.I.: How Weak Links Tame the Growth Explosion." Unpublished.
Kokotajlo, Daniel, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean. 2025. "AI 2027." AI Futures Project, April 3. https://ai-2027.com/.
Konopinski, Emil J., Cloyd Marvin, and Edward Teller. 1946. *Ignition of the Atmosphere with Nuclear Bombs*. Los Alamos National Laboratory Technical Report LA-602.
Korinek, Anton, and Donghyun Suh. 2024. "Scenarios for the Transition to AGI." NBER Working Paper 32255.
Kremer, Michael. 1993. "The O-Ring Theory of Economic Development." *Quarterly Journal of Economics* 108 (4): 551–76.
MacAskill, William. 2022. *What We Owe the Future*. Basic Books.
METR. 2026. "Task-Completion Time Horizons of Frontier AI Models." May 8. https://metr.org/time-horizons/.
Moore, Gordon E. 1965. "Cramming More Components onto Integrated Circuits." *Electronics* 38 (8): 114–117.
Mousa, Deena. 2025. "The Algorithm Will See You Now." *Works in Progress* 20, September 25. https://worksinprogress.co/issue/the-algorithm-will-see-you-now/.
Narayanan, Arvind, and Sayash Kapoor. 2025. *AI as Normal Technology*. Knight First Amendment Institute, April 2025.
Nielsen, Michael. 2024. "How to Be a Wise Optimist about Science and Technology?" December 1. https://michaelnotebook.com/optimism/index.html.
Nordhaus, William D. 2021. "Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth." *American Economic Journal: Macroeconomics* 13 (1): 299–332.
Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Book Group.
Posner, Richard A. 2004. *Catastrophe: Risk and Response*. Oxford University Press.
Rees, Martin. 2003. *Our Final Century: Will the Human Race Survive the Twenty-First Century*. William Heinemann.
Restrepo, Pascual. 2025. "We Won't Be Missed: Work and Growth in the AGI World." NBER Working Paper 34423.
Solow, Robert M. 1987. "We'd Better Watch Out." *New York Times Book Review*, July 12, 36.
Trammell, Philip, and Dwarkesh Patel. 2025. "Capital in the 22nd Century." December 29. https://philiptrammell.substack.com/p/capital-in-the-22nd-century.
US Bureau of Economic Analysis. 2025. "National Income and Product Accounts." US Department of Commerce. https://www.bea.gov/products/national-income-and-product-accounts.

US Department of Transportation. 2026. "Departmental Guidance on Valuation of a Statistical Life in Economic Analysis" https://www.transportation.gov/office-policy/transportation-policy/revised-departmental-guidance-on-valuation-of-a-statistical-life-in-economic-analysis (accessed January 10, 2025).

US Environmental Protection Agency. 2026. "Mortality Risk Valuation." https://www.epa.gov/environmental-economics/mortality-risk-valuation (accessed December 30, 2024).

Waymo. 2026. *Waymo Safety Impact*. https://waymo.com/safety/impact/ (accessed May 15, 2026).

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–45. Oxford University Press.

Zeira, Joseph. 1998. "Workers, Machines, and Economic Growth." *Quarterly Journal of Economics* 113 (4): 1091–117.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 23–46

The Emerging Market for Intelligence: How Firms Buy and Sell AI

Mert Demirer, Andrey Fradkin, and Nadav Tadelis

In the business-to-business market for large language model (LLM) inference, firms pay LLM providers to process their prompts and return output text in response. In late 2023, the price of state-of-the-art artificial intelligence in this market was roughly $30 per million tokens—with each token representing roughly three-quarters of a word of text. By late 2025, the same level of intelligence cost about three cents—a 1,000-fold decline. No other major technology has seen its quality-adjusted price fall so quickly.

Artificial intelligence is widely regarded as a general-purpose technology (Bresnahan and Trajtenberg 1995; Eloundou, Manning, Mishkin, and Rock 2024), one that, like electricity or the microprocessor, has the potential to reshape production across every sector of the economy. But general-purpose technologies do not diffuse instantly throughout the economy. They require institutions that set prices, match buyers to sellers, and transmit information about quality. The market for LLMs is one such institution, intermediating how AI capabilities are priced, evaluated, and deployed across the economy.

We describe the features of this market and document key empirical patterns in pricing, entry, usage, and market dynamism. Our analysis uses data from OpenRouter,

■ *Mert Demirer is Associate Professor of Economics, MIT Sloan School of Management, Cambridge, Massachusetts. Andrey Fradkin is Associate Professor of Marketing, Boston University Questrom School of Business, Boston, Massachusetts. While writing this paper, he was on leave at Amazon Inc. Nadav Tadelis is Senior Economist, Office of the Chief Economist, Microsoft Corporation, Seattle, Washington. Their email addresses are mdemirer@mit.edu, fradkin@bu.edu, and nadavtadelis@microsoft.com.* *The authors have received financial support from interested parties.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20261506.

a major intermediary in the LLM ecosystem that connects buyers and sellers. OpenRouter operates a standardized Application Programming Interface (API), a protocol that lets one program send requests to another and receive responses. OpenRouter provides access to hundreds of LLMs from dozens of providers, allowing us to observe pricing, entry, and usage across a broad cross-section of the business AI market from 2023 to 2025.

The supply side of this market has expanded rapidly. Our focus is on models readily available to firms for commercial use through APIs, a set that grew from 270 at the beginning of January 2025 to 668 by December 2025. Behind this proliferation, the number of firms developing these models has nearly doubled, while the number of inference providers—firms that host and serve models—has more than tripled over the same period. (The broader universe of artificial intelligence models is vastly larger still: the Hugging Face platform alone hosts millions of models.)

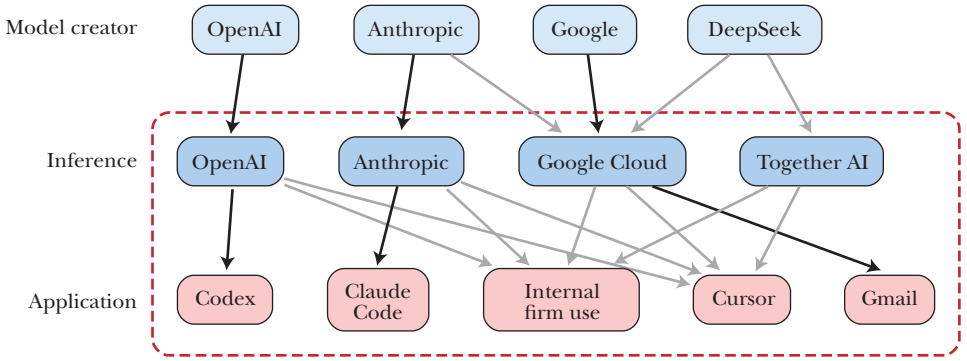
Competition in this market is intense. The leading model shifts frequently as successive releases displace incumbents at the capability frontier, reflecting rapid vertical innovation. The most intelligent model at any point in time does not win the majority of demand. There is substantial horizontal differentiation: no single model dominates across all use cases, and different models specialize in distinct tasks. These forces create a competitive environment characterized by both rapid obsolescence and specialization.

This proliferation of models and providers recalls similar episodes in economic history. For example, the US automobile industry had roughly 250 manufacturers by 1910, just a decade after the first commercial vehicles appeared; a shakeout then winnowed the field to a handful of survivors (Klepper 2002). The internet service provider market of the mid-1990s saw thousands of local ISPs spring up, only to consolidate as scale economies took hold (Greenstein 2015). We describe the ways in which the LLM market is now in a similar phase. Whether the market for intelligence follows a similar path toward concentration or instead fragments into specialized niches will depend on the strength of scale economies, the pace of capability improvements, and interactions across the industry's vertical layers.

Our focus is on business adoption of large language models. We do not study consumer use through chat interfaces (Chatterji et al. 2025; Bick, Blandin, and Deming 2024), or the broader labor market (Eloundou, Manning, Mishkin, and Rock 2024; Brynjolfsson, Li, and Raymond 2025; Acemoglu and Restrepo 2019) and macroeconomic (Acemoglu 2025) consequences of AI. These are important and rapidly growing areas of inquiry. What LLM markets offer is a unique window into how firms are accessing intelligence as a productive input.[1]

[1] The numbers and figures in this paper reflect the state of the market as of December 2025, and the text was last updated in April 2026. This market is moving rapidly, and some specifics may be out of date by publication.

Figure 1
Market Structure: Model Layer, Inference Layer, and Application Layer (Illustrative Firms in Each Layer)



Source: Authors' creation.
Note: The figure illustrates the vertical structure of the market for LLM inference. The Model layer comprises the frontier labs that train and maintain foundation models. The Inference layer consists of cloud and specialized providers that host models and process API requests. The Application layer includes labs' own products, third-party applications built on top of inference APIs (for example, Cursor and Perplexity), and internal firm use of LLMs. Bold arrows indicate vertical integration: Google owns Google Cloud and Gmail; OpenAI owns Codex; Anthropic owns Claude Code. Thin arrows indicate market-based supply relationships between independent firms. The dashed box marks the focus of this paper: the interface between inference providers and customers.

How the Market Works

The market for LLM inference is organized as a vertical supply chain with three layers: creators who build models (the model layer), inference providers who host and serve them (the inference layer), and end consumers—firms that either integrate LLM capabilities into their internal workflows or embed them in downstream applications that serve their own customers (the application layer). Figure 1 offers an illustration. Some players occupy multiple positions in the supply chain: all major model creators both train models and serve inference through their own platforms, and several—such as Anthropic and OpenAI—extend further into the application layer by offering their own end-user products, such as coding assistants. In this paper, we focus on the interface between inference providers and customers.

The Supply Chain

Model creators are the firms that design and train LLMs. Training a frontier model requires assembling massive datasets, designing a model architecture, and executing computationally intensive training runs that can cost hundreds

of millions of dollars (Emberson and Edelman 2025). The leading closed-source creators—Anthropic, Google, and OpenAI—keep their model weights proprietary. For example, Google develops models internally and offers them directly through its cloud platform. In some cases, closed-source creators distribute models through multiple providers to expand reach—Anthropic, for instance, makes its Claude models available through AWS, Microsoft Azure, and Google Cloud in addition to its own platform.

Other creators, most notably Meta and the Chinese firm DeepSeek, release their model weights under open-source (or open-weight) licenses, allowing anyone to download and run the model. The distinction between open-source and closed-source models is fundamental to the market's competitive structure (Lerner and Tirole 2005). Closed-source models are available only through the creator or its authorized partners. Open-source models, by contrast, can be hosted by any firm with sufficient computing capacity. They can also be modified: firms routinely fine-tune open-source models for specific tasks or "distill" large models into smaller, faster variants that sacrifice some capability for lower cost.

Inference providers are the firms that actually run models in response to user requests. When a user sends a prompt to an API, it is the inference provider's servers that process the input and return the model's output. We can categorize inference providers into three groups. The first group consists of the model creators themselves. As discussed above, most major model creators operate their own inference services and host their models directly for customers. The second group comprises major cloud platforms, such as Amazon Web Services and Google Cloud. These platforms provide inference services for both open- and closed-source models, with closed-source models often subject to exclusive or preferential access arrangements. Little is known about the structure and terms of these contracts. The third category is a growing set of specialized inference companies—firms like Together AI, Cerebras, and Groq—that focus on serving open-source models. These firms compete by optimizing hardware and software to deliver faster response times (lower latency), higher processing speeds (greater throughput), and lower prices. The result is that popular open-source models are often available from a dozen or more competing providers, each offering different combinations of price, speed, and reliability. This stands in sharp contrast to closed-source models, which are typically available from just one or a few providers.

Inference providers, therefore, compete along different dimensions depending on the type of model they serve. For providers serving closed-source models, the model itself is a key source of differentiation. By contrast, for providers serving open-source models, competition shifts to governance and the quality of inference, including latency, throughput, reliability, and price. For major cloud providers, existing customer relationships create an additional source of competitive advantage. Firms whose data, infrastructure, and workflows are already embedded within a particular cloud ecosystem may face switching costs, making integration, security compliance, and bundled service offerings important dimensions of competition.

Customers are primarily enterprises that purchase API access to incorporate AI into their operations or products.[2] Enterprise usage broadly falls into two categories. The first is internal use: a firm uses OpenRouter or directly goes to an inference provider for its own workflows, for example, to automate document review, generate code, or power internal search. The second is the application layer: firms build products in which the LLM is a core component embedded in a customer-facing service, such as a legal-tech tool that analyzes contracts or a coding assistant that offers code completion. These LLM-powered applications can offer a single model or a portfolio of models to their users; when multiple models are available, the end user often selects the model. Across both categories, model choice reflects the nature of the underlying task: cost-sensitive, high-volume workflows such as those used for e-commerce search gravitate toward cheaper and faster models, while high-stakes and more complex workflows such as vulnerability detection in software gravitate toward expensive frontier models.

The application layer is highly dynamic. A large ecosystem of startups and incumbent firms is developing AI-powered applications across a wide range of categories, including accounting, law, medicine, marketing, customer support, and education. The most mature segment to date is coding, where developer-assistant tools have achieved substantial adoption. The best-known third-party tools are GitHub Copilot and Cursor, which integrate access to many frontier models within a coding assistant interface. These compete with vertically integrated offerings, such as Codex from OpenAI and Claude Code from Anthropic, which serve their own models. Applications in this layer differ not only in the models they provide access to, but also in how they are designed, how well they fit into a user's existing tools and processes, what features they offer to larger organizations (such as security controls and team administration), how they price their service, and the quality of their customer support. As a result, competition at the application layer involves both model performance and nonmodel product differentiation.

For most enterprises, entry into this ecosystem begins by signing up directly with an inference provider and accessing models through standardized API contracts with usage-based pricing. At this scale, relationships are typically governed by online terms of service, rather than bespoke agreements. For high-volume users, however, procurement more closely resembles traditional enterprise information technology contracting: large firms often negotiate individualized agreements with inference providers that specify volume discounts, committed capacity (a guaranteed amount of computing reserved for the buyer), service-level guarantees (commitments on uptime and response speed), data governance provisions, and security compliance requirements.[3]

[2] Individual users may also create accounts and access these APIs directly on most platforms. However, such usage is relatively uncommon, as most consumer-facing products operate on a subscription basis.

[3] In addition, a separate enterprise-services segment has emerged in which AI firms—or their implementation partners—work with clients to fine-tune models, build firm-specific AI agents, and integrate systems into proprietary workflows. In such cases, the enterprise may contract directly with the

While creating an account with any single inference provider is straightforward, managing accounts across multiple providers and customizing queries for each can be operationally costly for firms that wish to compare prices, optimize performance, or switch across models. To address this problem, a fourth layer has recently emerged: the aggregator layer, which sits between the inference layer and consumers and provides a single standardized API that enables access to models from multiple providers. Beyond consolidating access, aggregators can dynamically route requests to the lowest-cost or fastest provider of a given model and offer automatic fallback during outages, thereby reducing switching frictions and intensifying competition within the inference layer. The leading aggregator is OpenRouter, which offers access to hundreds of models from dozens of providers. Although this segment is expanding rapidly, most tokens are still processed directly by inference providers rather than through aggregators.

Tokens, Pricing, and How Firms Pay

The fundamental unit of transaction in this market is the token. When a firm sends a request to a large language model through an API, it sends a sequence of prompt tokens (the input), and the model returns a sequence of completion tokens (the output). Providers charge per million tokens, with separate rates for prompt and completion tokens. Completion tokens are typically more expensive, because generating output is more computationally intensive than processing input. This token-based pricing model means that AI inference is purchased much like a utility—firms pay for what they use, with no fixed costs for hardware or model deployment. Costs scale linearly with usage: a firm that doubles the number of documents it processes simply pays for twice as many tokens.

Inference providers post their prices, and most customers are price takers. Supply and demand are balanced primarily through nonprice mechanisms. The primary channel is quality degradation: when demand exceeds capacity, latency rises and throughput falls, effectively rationing access by reducing service quality rather than raising prices. A second mechanism is batch processing, in which users submit requests and the provider processes them within 24 hours at a discounted rate, shifting demand to off-peak capacity. While prices can and do change over time, real-time price adjustments are not widely used to balance supply and demand.

Pricing varies enormously across models and providers. Frontier closed-source models—the most capable systems available—may charge $15 or more per million completion tokens, while commodity open-source models can cost less than a cent for the same volume. This variation reflects several factors: the model's intelligence, whether it is open- or closed-source, the provider's speed and reliability, and the competitive dynamics of the specific market segment.

In addition to prompt and completion tokens, many models feature "reasoning tokens"—tokens required for the model's internal deliberation steps—and "cache

model provider or with a third-party integrator that builds and maintains customized applications on top of underlying API access.

tokens," which allow users to reuse frequently sent prompts at a discount. These pricing primitives create a schedule in which firms can optimize their costs by choosing models, providers, and prompt strategies that match their particular usage patterns.

To make the pricing concrete, consider an example of a lawyer who submits a 50-page contract to Claude Opus 4.6 for a one-page summary. The text in the contract constitutes the prompt tokens, and the model's summary output constitutes the completion tokens. Assuming roughly 500 tokens per page, the input tokens cost about 13 cents (25,000 tokens at $5 per million) and the output tokens cost about one cent (500 tokens at $25 per million). Because Opus 4.6 is a reasoning model, the lawyer also pays for reasoning tokens, which cost $25 per million tokens, though the number of reasoning tokens used is less predictable and varies by model. Now suppose the lawyer follows up with a question about liability exposure across multiple clauses. At this point, the original contract is already stored in the model's "context cache," which is a temporary memory that stores recently submitted text, allowing the model to reference it without reprocessing. The lawyer pays a discounted rate for the cached tokens (25,000 tokens at $0.50 per million), a standard prompt rate for the text of the follow-up question, and completion tokens for the response itself.

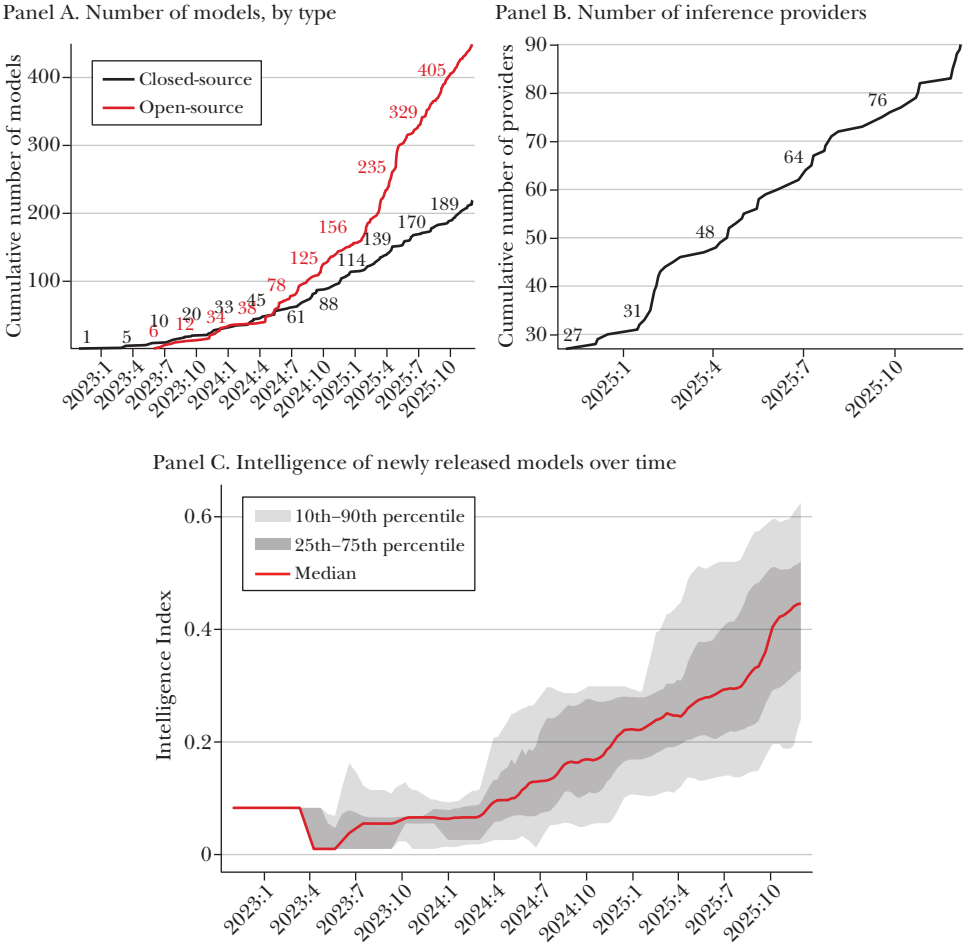
Data Sources

We study this market using data from OpenRouter, the leading LLM aggregator. OpenRouter provides a standardized API for accessing hundreds of models from dozens of providers, and its usage data allow us to observe token consumption by model, date, and by application category (for example, programming, marketing, and translation). We also observe pricing data for each model and provider, as well as information on model characteristics such as intelligence benchmarks, latency, and throughput. OpenRouter data are available from July 2023, with a balanced daily panel from January 2025 through December 2025.[4]

OpenRouter provides API access to virtually all commercially relevant LLMs and introduces new models immediately after creators release them. The prices observed in the OpenRouter data reflect the actual posted API prices charged by providers. As a result, the data provide an accurate view of the evolving supply side of the market—pricing, entry, and model turnover—without requiring us to reconstruct these features from multiple proprietary sources. On the demand side, we observe the aggregate usage of each model by provider at the daily level. Because OpenRouter accounts for only a small share of total API usage and skews toward developers and technology firms, we do not interpret our results on model demand as estimates of aggregate market shares, but instead use these data to study relative differences across models, use cases, and changes over time.

[4]We obtained these data via a scrape of the website, and are making it available as part of the replication package (Demirer, Fradkin, and Tadelis 2026).

Figure 2
Rapid Growth in Models, Providers, and Intelligence



Source: OpenRouter and Artificial Analysis.
Note: Panel A shows the cumulative number of models available on OpenRouter, separated by open- and closed-source. Panel B shows the cumulative number of inference providers. Panel C shows the distribution of the Intelligence Index for models released within six months of each date. The red line is the median; shaded areas show the interquartile and 10th–90th percentile ranges.

The Explosion of Supply

The Explosion of Models and Creators

Figure 2 summarizes the growth of the market along three dimensions: models, providers, and intelligence. Panel A shows the cumulative number of models available through OpenRouter over time, separated into open-source and closed-source categories. In early 2024, 63 models were available. By December 2025, that

number increased to 668. The expansion is driven overwhelmingly by open-source models: as of December 2025, there are 449 open-source models compared to 219 closed-source models. Importantly, this growth reflects not just a few prolific creators releasing many models: the number of distinct model creators has nearly doubled, from 47 at the start of 2025 to 88 by December, with the vast majority being open-source creators (78 open-source versus 10 closed-source).[5] The number of closed-source creators has been stable at 10 since April 2025.

This proliferation of open-source models has important economic implications. In traditional technology markets, high fixed costs and proprietary standards tend to limit entry. The open-source model—in which creators bear the training costs but make weights freely available—dramatically lowers the barrier to downstream competition. Meta invested billions in training its Llama models, and then gave them away (Nagle and Yue 2025). A Chinese model creator, DeepSeek, operating in a different regulatory environment, does the same. Any firm with GPUs (graphics processing units—specialized computer chips used for AI computation) can begin serving an open-source model, and anyone with machine-learning expertise can fine-tune or distill an existing model for a specialized application.

A key question is whether it is worthwhile for firms to create open-source models, given the presence of external inference providers that compete to monetize them. Several reasons have been given for open-sourcing, including that open-source models are complements to algorithmic recommendations (for example, for Meta) and hardware (for example, for Nvidia), and that open-source models hurt competitors with closed-source models. These open-source releases function as a competitive subsidy to the rest of the market, keeping prices and entry barriers low in a way that has little parallel in other capital-intensive technology industries. The result is an ecosystem in which innovation occurs not only at the frontier but also in the adaptation and deployment of existing capabilities.

The Rise of Inference Providers

Panel B of Figure 2 documents the rapid entry of inference providers. In November 2024, there were 27 providers on OpenRouter. By December 2025, that number had grown to 90. This entry is concentrated in the provision of open-source models: popular open-source models are often hosted by a dozen or more competing providers. In contrast, closed-source model creators often serve their models from just a few providers, with new providers typically gaining access through investment partnerships with the creator.

When multiple providers offer the same underlying model, they can differentiate on other dimensions: price, latency, and reliability. Our data show substantial variation across these dimensions. For example, among providers hosting the same open-source model, throughput can vary by a factor of five, and prices can

[5]A "creator" is the organization that publishes the model, identified by OpenRouter's publisher field (for example, Anthropic, Google, Meta, DeepSeek). This corresponds to the corporate brand releasing the model rather than a legal taxpayer entity, and generally to the corporate parent.

differ by more than an order of magnitude (Demirer, Fradkin, Tadelis, and Peng 2025). These seemingly homogeneous products (the same model weights) are sold by differentiated sellers, with price and nonprice competition. The contrast with closed-source models is stark: their prices remain largely rigid over their lifetimes and there is almost no variation in price across providers as model creators can dictate prices through contractual arrangements (Demirer, Fradkin, Tadelis, and Peng 2025).

Measuring Capabilities

A central challenge in the market for LLM inference is measuring differences in model quality. To address this challenge, the industry has developed benchmarks—standardized evaluation suites that test models on tasks such as mathematical reasoning, coding, and scientific problem solving and assign quantitative performance scores. These benchmarks are typically highly correlated across domains, suggesting a dominant general capability factor rather than isolated task-specific strengths (Papailiopoulos 2026). This general capability may stem from model or training-data size, consistent with well-documented scaling laws (Kaplan et al. 2020).

While benchmarks are useful, they are imperfect because real-world enterprise tasks may differ from exam-style evaluations. In what follows, when we discuss model capabilities, we measure them using the composite Intelligence Index reported by Artificial Analysis, which is an independent evaluation platform that benchmarks AI models across multiple dimensions. Its Intelligence Index combines ten standardized evaluations spanning reasoning, knowledge, mathematics, and coding. It is important to note that other model characteristics such as latency, modality (image, audio, or text), and cost may determine which model is used for a particular task, especially as some tasks do not require frontier intelligence.

Improving Intelligence

Panel C of Figure 2 tracks the capabilities of newly released models over time, where "new" is defined as released within the preceding six months. The red line plots the median intelligence of new releases, while the shaded regions depict the 25th–75th and 10th–90th percentile ranges. The median intelligence of new models has risen from about 0.1 when GPT-3.5 launched to above 0.4 today—roughly a fourfold improvement. This represents a shift from models that frequently hallucinate and struggle with multi-step reasoning to models that can solve complex problems and have much lower rates of hallucination (Kalai, Nachum, Vempala, and Zhang 2025). The distribution of intelligence across newly released models has also widened: the best models today score roughly six times higher than the earliest models in our sample, yet firms continue to release models with lower capabilities as well. This improvement in capabilities is a supply-side story with demand-side consequences. As models become more capable, they become useful for a wider range of tasks, potentially expanding the total addressable market for AI services. At the same time, the rapid pace of improvement creates strong incentives for firms to stay near the frontier.

Taken together, the evidence indicates that the industry is in an explosive expansion phase. Entry occurs simultaneously across multiple layers of the supply chain—model creation, inference provision, and application development—while rapid improvements in capabilities continually reshape the competitive frontier. Open-source releases play a central enabling role, lowering downstream entry barriers and accelerating experimentation. The result is a layered and quickly changing ecosystem in which specialization, vertical integration, and horizontal differentiation coexist, and in which competitive conditions can shift quickly as new models and providers enter.

The Price of Intelligence

Perhaps the single most striking fact about the business-to-business AI market is the speed at which the price of intelligence has fallen. To be sure, plunging prices in some form have been seen with other technologies. The deregulation of telecommunications in the 1990s cut the price of a long-distance phone call by about 90 percent over 15 years. Gordon Moore, who would later be a cofounder of Intel, predicted back in 1965 when working for Fairchild Semiconductor that the number of transistors on a chip would double every two years (Moore 1965). Moore's Law delivered roughly a hundredfold reduction in the cost of computing per decade. The LLM market has compressed an even larger price decline, which might fairly be characterized as a price collapse, into just two years.

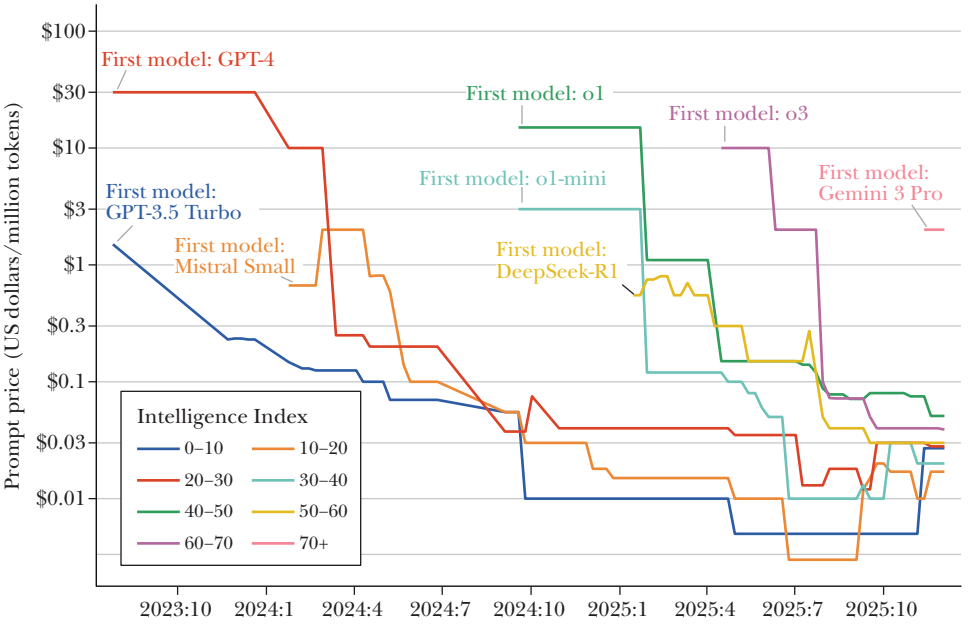
The 1,000-Fold Price Decline

Figure 3 shows the minimum price for models at each level of intelligence (the vertical price axis is a log scale). Each line represents a distinct capability tier, measured by the Intelligence Index. In mid-2023, GPT-4-class intelligence cost roughly $30 per million prompt tokens—the only model at that capability level. By late 2025, the same level of intelligence is available for about three cents, a decline of approximately 1,000-fold over two years. The pattern is not limited to GPT-4-class models: steep price declines are visible across all intelligence tiers.

Three forces drive these declines. First, advances in model architecture and training techniques allow creators to achieve the same level of intelligence with fewer neural network parameters. This translates into less computation required per query. Second, improvements in hardware efficiency and serving infrastructure enable inference providers to deliver a given model at lower marginal cost, reducing the price of inference even while holding model capability constant. Finally, competition—both between closed-source providers and from new open-source models—puts downward pressure on prices.

The implications of this price decline for firm adoption are profound. Tasks that were prohibitively expensive to automate a year ago are now economically viable, and the set of applications for which AI is cost-effective is expanding rapidly. At the same time, firms shift token consumption to more intelligent models, so that the average

Figure 3
The Falling Price of Intelligence



Source: OpenRouter and Artificial Analysis.
Note: This figure shows the minimum prompt price per million tokens for models at each level of intelligence over time. Each line represents a distinct bin of intelligence scores. The y-axis uses a log scale.

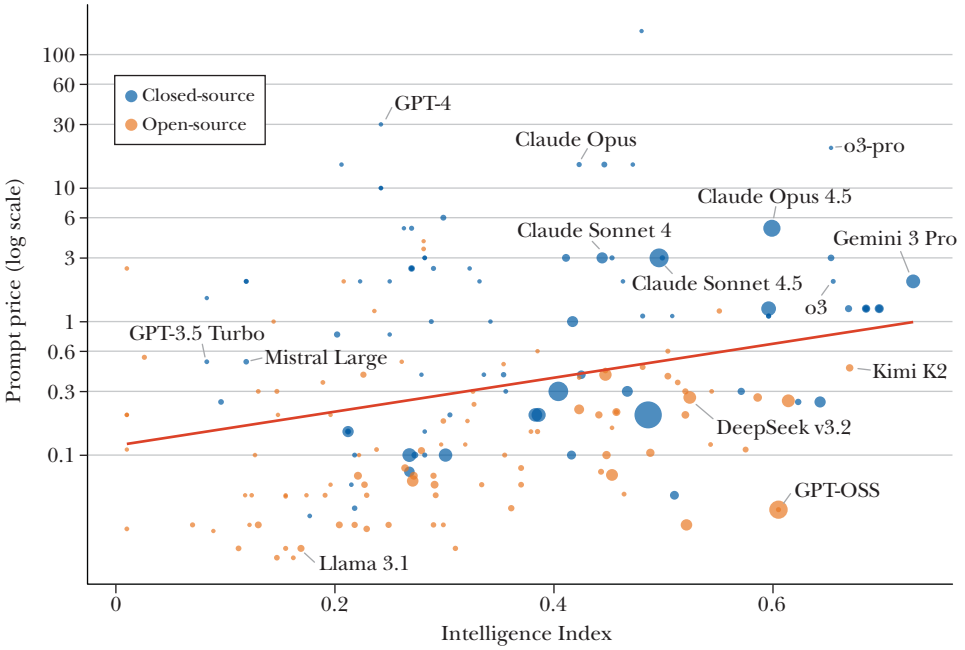
per-token price paid remains relatively stable (Demirer, Fradkin, Tadelis, and Peng 2025). Firms are not simply buying the same intelligence more cheaply—they are upgrading to more capable models as prices fall. The intelligence of models in use has risen steadily, with the token-weighted median intelligence score increasing from 0.30 in January 2025 to 0.44 by year-end (Demirer, Fradkin, Tadelis, and Peng 2025). This pattern is consistent with firms using price declines to purchase better, not just cheaper, AI. The dynamic echoes the history of personal computers and smartphones—modest price declines paired with order-of-magnitude capability gains—except that for LLMs, both prices and capabilities have moved sharply in the buyer's favor.

The Price-Intelligence Relationship

Figure 4 shows the relationship between price and intelligence across all models as of December 2025. Each point is a model; color indicates whether the model is open-source (orange) or closed-source (blue); and point size reflects total token usage.

Two patterns stand out. First, the relationship between price and intelligence is approximately linear in logs, meaning that it is highly nonlinear in levels. Moving from a mid-tier to a frontier model is associated with a disproportionately large

Figure 4
Price Versus Intelligence, Open-Source Versus Closed-Source



Source: OpenRouter and Artificial Analysis.
Note: Scatter plot of prompt price (log scale) against the Intelligence Index as of December 6, 2025. Point size is proportional to total token usage. Color indicates open-source (orange) versus closed-source (blue) models. The fitted line shows the average price-intelligence relationship.

increase in cost. This is consistent with the economic intuition that producing the last increment of intelligence is far more expensive than producing the first. Second, there is enormous price dispersion at any given level of intelligence. For models with similar benchmark scores, prices can vary by more than two orders of magnitude.

Open-Source versus Closed-Source Models

Much of the price dispersion turns on a fundamental distinction in how models are distributed. Closed-source models, such as OpenAI's GPT series and Anthropic's Claude, are proprietary: the creator keeps the underlying model weights secret and sells access only through its own API or a small number of authorized partners. Because the creator controls the sole channel to the model, it can set the price directly, and competition for that specific model is limited. Open-source models, such as Meta's Llama and DeepSeek's releases, are distributed differently: the creator publishes the model weights, allowing any firm with sufficient computing capacity to host and serve the model. This means a single open-source model is typically offered by many competing inference providers, and price competition among them drives the cost of serving that model down toward marginal cost.

Controlling for intelligence level, open-source models are approximately 90 percent cheaper than their closed-source counterparts. Despite the price advantages, the open-source models are used far less than their closed-source counterparts, as indicated by the point sizes in Figure 4, which scale with token volume. Overall, open-source models account for less than 30 percent of total token consumption.

Why would firms pay substantially more for closed-source models when open-source alternatives with comparable benchmark performance are available at a fraction of the cost? Several explanations are plausible, and they are not mutually exclusive.

First, benchmarks for performance may not fully capture the dimensions of quality that firms care about. Closed-source creators invest heavily in "reinforcement learning from human feedback" (often abbreviated as RLHF), a training process in which human evaluators rank model outputs to iteratively improve response quality. They also invest in "safety alignment," which shapes how a model reflects human values and ethics. These post-training refinements may improve the model's usefulness for real-world business tasks in ways that conventional performance benchmarks may not measure. Second, closed-source models come with contractual guarantees, enterprise support, and service-level agreements that reduce the risk to firms integrating AI into critical business processes. Third, adjustment costs may play a role to the extent that a firm that has optimized its prompts, workflows, and internal tooling around a particular model faces costs in migrating to an alternative, even if the alternative is cheaper. Fourth, firms may be to some extent mistaken in perceiving open-source models as inferior substitutes for closed-source alternatives, even when actual performance is similar—a belief that, whether justified or not, can sustain the price premium.

The coexistence of expensive closed-source and cheap open-source models is one of the most economically interesting features of this market. It suggests that the market for AI services is not a simple commodity market in which the lowest-cost producer wins. Instead, it is a differentiated market in which firms trade off multiple attributes—intelligence, price, reliability, brand trust, and integration costs—when choosing a model.

Market Dynamics and Differentiation

So far, the business-to-business market for artificial intelligence resembles a race in which some creators compete to advance the frontier of intelligence and charge a premium, while other creators offer more targeted models optimized for specific tasks. Over time, one possibility given the number of models and falling prices is that this market results in a race to the bottom, with intelligence becoming a commodity priced close to marginal cost. However, this outcome has not happened yet.

An important aspect of this market is heterogeneous demand across use cases. A software firm generating code, a law firm summarizing depositions, and a retailer

translating product listings face different task requirements, latency constraints, reliability needs, and willingness to pay. As a result, technological progress does not collapse the market into a single dominant product.

A Market in Motion

To illustrate these dynamics, we divide the market into a dozen use cases: trivia, translation, technology, science, roleplay, programming, search engine optimization (SEO), marketing, legal, health, finance, and academia. Figure 5 makes these dynamics visible. Each row corresponds to a use-case category, and the horizontal axis spans the weeks in 2025. At each point in time, the color indicates which creator's model holds the largest share of usage in that category, while the labels within each segment identify the specific model leading during that period. We highlight the top six creators—Google, Anthropic, OpenAI, DeepSeek, xAI, and Meta—and group all others into an "Other" label.

Two patterns stand out. First, no single creator dominates across all categories: the figure is a mosaic of colors rather than a monochrome. Second, the dominant color within any row shifts frequently as new model releases displace incumbents. The leading model in the legal category changed multiple times over the year; translation, marketing, and finance followed similarly volatile patterns.

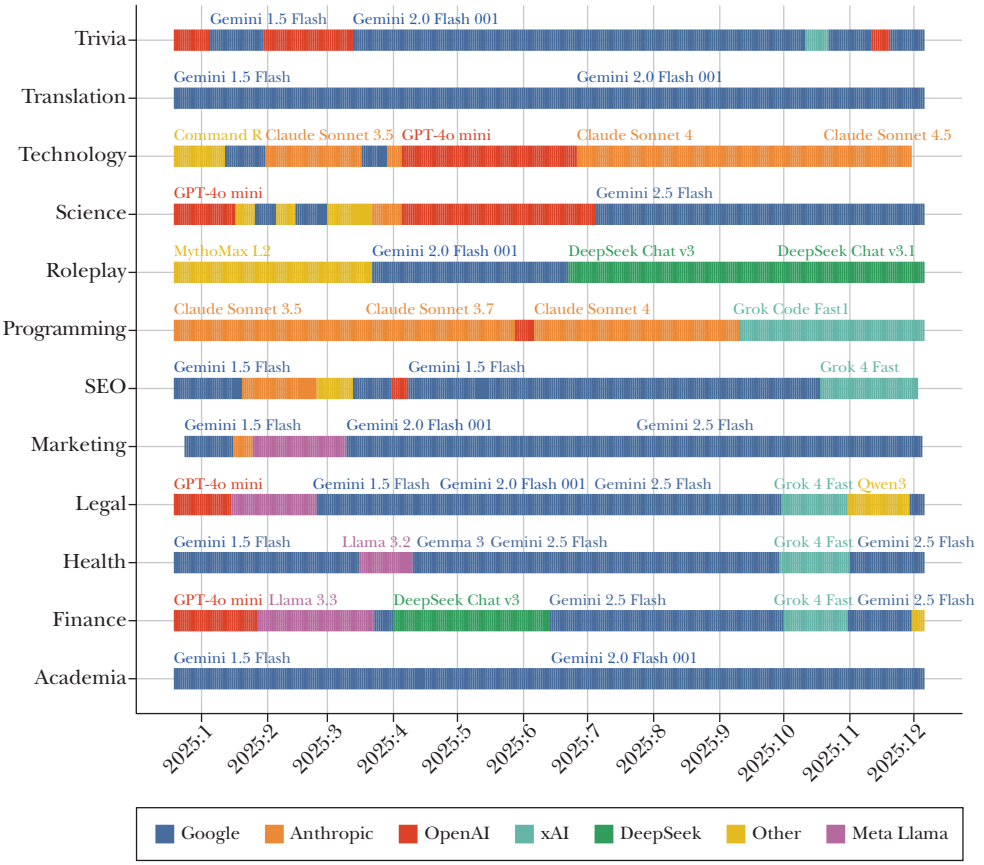
The churn is even more pronounced at the model level. Over this period, several major model launches occurred, and across categories, leading models were frequently replaced by newly released versions. In programming, for example, the leading model moved quickly from Claude Sonnet 3.5 to 3.7, and then to 4 and 4.5, closely following their release dates. Similarly, in categories where Google leads, we observe transitions toward newer frontier models. Importantly, adoption is not uniform across use cases. Some categories (such as programming and legal) transition rapidly to newly released models, whereas others (such as trivia and marketing) continue to rely on earlier versions.

Horizontal Differentiation

The cross-category patterns of which model leads also reveal horizontal differentiation in addition to the vertical innovation described above. Different use cases require different combinations of capability, speed, reliability, and price. A developer using a coding assistant needs a model that can reason through complex logic and generate correct code—errors are immediately apparent when the program fails to run. A marketing team generating ad copy places greater weight on tone, creativity, and stylistic control. A customer-service chatbot must prioritize speed and low cost, as it processes large volumes of relatively low-value queries. These differing requirements steer buyers toward different regions of the capability-price frontier, creating persistent segmentation.[6]

[6]These patterns are also visible beyond OpenRouter. For example, Anthropic's Claude models are known to be especially popular among software developers: coding tasks account for the largest share of usage on Anthropic's own platform (Handa et al. 2025), and enterprise surveys rank Claude models first in code generation (Tully, Redfern, Das, and Xiao 2025).

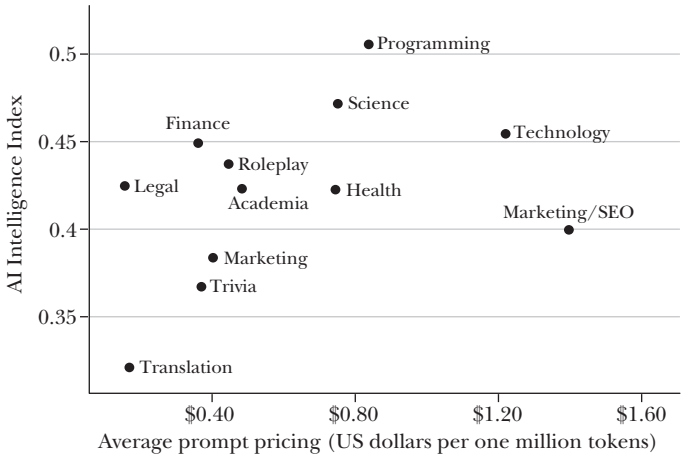
Figure 5
Market Leadership Shifts Across Use Cases



Source: OpenRouter.
Note: Each row is a use-case category. The underlying data are weekly: each segment within a row corresponds to a week in 2025, while the horizontal axis is labeled by month. Color indicates the creator of the model holding the largest share of prompt tokens in that category during that week (30-day rolling window). Frequent color changes within rows reflect the leading model changing rapidly following new releases.

Figure 6 documents this phenomenon by plotting the average AI Intelligence Index (yaxis) and prompt pricing (x-axis) by category. Programming is the most intelligence-intensive application, with an average score of 0.51, followed by science at 0.47. At the other end, translation (0.32) and trivia (0.37) rely on models with substantially lower intelligence scores. The variation in price paid across categories is substantial: search engine optimization applications pay an average of $1.40 per million prompt tokens, while legal and translation applications pay just $0.16 and $0.17, respectively—a nearly ninefold difference. Applications that demand higher

Figure 6
Intelligence and Pricing Vary Across Use Cases



Source: OpenRouter and Artificial Analysis.
Note: The figure shows the usage-weighted average intelligence score and prompt price per million tokens used in each category over a 30-day period (November 7 to December 6, 2025).

capability tend to pay more, while high-volume or simpler tasks gravitate toward lower-priced models.

This interaction between vertical and horizontal differentiation helps explain several features of the market. It explains why many creators can coexist: a firm that produces a lower-intelligence model at a low price is not simply losing to more capable competitors—it is serving a different segment of demand. It explains why frontier models, despite their impressive capabilities, account for only a modest share of total token consumption (as illustrated earlier in Figure 4): for many applications, the incremental intelligence does not (yet) justify the higher cost. And it explains why the market has not converged on a single winning model, despite the fact that one model is measurably "best" on standardized benchmarks at any given moment.

These results have two important implications. First, model releases function as discrete competitive shocks that can rapidly reallocate demand. Because code-driven switching costs are minimal for most applications—an API call can be redirected without new infrastructure investment—users adjust quickly when relative performance shifts. Whether users do switch also depends on whether their workflows are generic rather than being optimized for only one model. Second, they imply that any competitive advantage derived from having the best model at a given moment is inherently temporary. The returns to innovation are fleeting, creating strong incentives for continuous investment in model improvement.

The dynamism we document is not without precedent. Early technology markets routinely exhibit turbulent leadership before consolidation sets in. In the early American automobile industry, hundreds of manufacturers competed in the first

two decades of the twentieth century, with market leadership shifting frequently as engine designs, body styles, and production methods evolved rapidly; by the 1930s, the industry had consolidated around a handful of dominant firms (Klepper 1996). The early search engine market followed a similar pattern: AltaVista, Lycos, Excite, and Yahoo all held leading positions in the late 1990s before a superior entrant displaced them within a few years. In both cases, the early phase was characterized by rapid entry, frequent leadership changes, and users' quick adoption of better alternatives—precisely the pattern we observe in LLMs today.

What is distinctive about the LLM market is the pace: the turnover we document occurs over months rather than years, reflecting both the speed of model development and the small cost of switching models via standardized APIs. The market's segmentation across use cases, intelligence levels, and price points also suggests that, if consolidation occurs, it is unlikely to yield a single dominant model in the enterprise LLM market. Different buyers have distinct needs, and different creators have identified niches to serve them. Whether the LLM market will eventually consolidate into a small number of dominant platforms, as the automobile and search markets did, or sustain a more fragmented equilibrium, remains an open question.

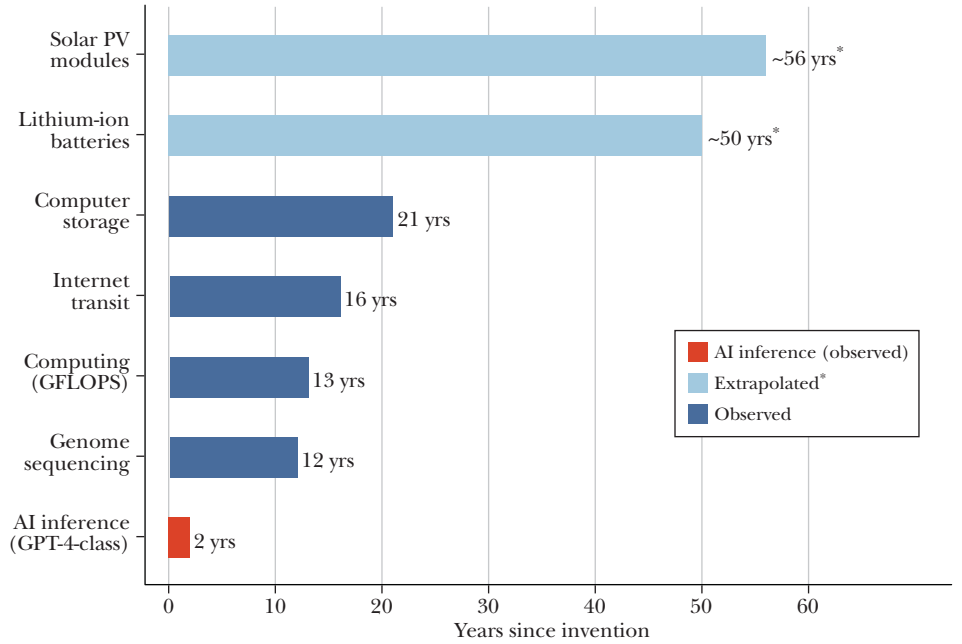
AI Adoption and Cost in Historical Perspective

The economic literature on general-purpose technologies documents a recurring pattern: widespread adoption and productivity effects arrive slowly, often decades after invention (Bresnahan and Trajtenberg 1995). Steam power and electrification took 40 years or more to diffuse broadly across the economy; computers and information technology took roughly 25 years from early commercial availability in the 1970s to pervasive business adoption in the mid-1990s.

To understand why these lags occur and whether the trajectory of AI technology might be different, it is useful to decompose adoption costs into two categories. The first is the *cost of access and usage*—what firms must pay before and while using the technology. Electrification required connection to a power grid that took decades to build, and electricity itself remained expensive for decades (David 1990). The internet required networking infrastructure, hardware, and technical expertise. Even after firms made the necessary investments to gain access to these new technologies, high usage costs limited the range of applications that were economically viable. Technologies diffuse widely only after both barriers fall: the technology must be accessible, and usage must be affordable.

The second category is *organizational adjustment costs*—the softer, often slower process by which firms and workers learn to use a new technology effectively. Even after electricity was widely available and affordable, factories took a generation to reorganize their floor layouts around distributed electric motors, rather than central steam shafts (David 1990). The productivity gains from information technology similarly lagged adoption by a decade or more, as firms redesigned workflows, retrained

Figure 7
Time for a 1,000-Fold Fall in Price



Source: OpenRouter data (Aubakirova et al. 2025); NHGRI genome sequencing costs via Our World in Data (National Human Genome Research Institute 2022; Roser, Ritchie, and Mathieu 2023); cost per GFLOPS (billions of floating-point operations per second, a measure of computing speed) (AI Impacts 2015, 2017); internet transit prices (DrPeering 2016); computer storage prices (Mathieu 2024); solar photovoltaic module prices (Ritchie, Rosado, and Roser 2023); lithium-ion battery prices (Ritchie and Rosado 2026).
Note: Each bar shows the number of years required for a technology's price to fall by a factor of 1,000. Solid bars denote technologies for which a 1,000-fold decline has been observed; light bars marked with * show extrapolations at the observed constant rate of decline for technologies that have not yet reached the 1,000-fold threshold.

workers, and developed new managerial practices (Brynjolfsson and Hitt 2000). These organizational co-invention costs are more difficult to observe and often slower to resolve, as they depend on firm-level capabilities, managerial practices, and industry-specific constraints that vary across sectors and use cases. Artificial intelligence, delivered through APIs, faces a fundamentally different cost structure. The physical infrastructure required for access—the internet, cloud computing platforms, electricity grids, and connected devices—already exists. A firm that wants to adopt LLMs does not need to build roads or string wires; it needs an API key and a few lines of code. This means that access to AI can diffuse far faster than access to electricity or the internet.

The cost of usage has also collapsed at an unprecedented rate. Figure 7 compares how quickly the prices of several foundational technologies declined, measured as the number of years required for a 1,000-fold fall in price. AI inference

stands out as an extreme case: as noted earlier, the price of GPT-4-class inference fell roughly 1,000-fold in about two years, far faster than internet transit, genome sequencing, computing, or computer storage, and vastly faster than the trajectories implied by batteries and solar. The broader pattern is that newer digital technologies can experience extraordinarily rapid cost compression. For AI, this means that both components of the cost barrier—the cost of access and the cost of usage—have fallen faster than for any comparable technology.

The low costs have encouraged rapid early adoption. A survey conducted in early 2026 reports that 43 percent of US workers have used AI in their jobs (Bick et al. 2026). If we look at country-level adoption, which is commonly used to study technology diffusion (as in Comin and Hobijn 2010), LLMs have already diffused: our data indicate usage in more than 200 countries and territories. No previous general-purpose technology has achieved this breadth of access so quickly.

However, the category of organizational adjustment costs remains, and early evidence suggests it may now be the binding constraint on realizing the economic potential of AI technologies. The same survey that documents widespread worker-level adoption finds that productive integration—using AI in the actual production of goods and services rather than for experimentation or personal assistance—lags substantially. Only 7 percent of US firms and 4 percent of European firms report using AI in production (Bick et al. 2026).[7] The gap between diffuse worker-level adoption and narrow firm-level production use suggests that the binding constraint on AI's economic impact has shifted from access and cost to organizational transformation.

The level of organizational investment in adapting to the possibilities of AI technology is sharply heterogeneous across industries. In software development, where work is already conducted through code in standardized environments—the same interfaces that LLMs use—integration costs are low, and productivity gains have already materialized: studies of AI-assisted coding find that developers complete tasks substantially faster and produce more code output (Peng, Kalliamvakou, Cihon, and Demirer 2023; Sarkar 2025). In other settings—healthcare, law, finance, manufacturing—the organizational bottleneck is considerably more binding. A hospital system considering the possibility of adopting AI diagnostic tools, for example, must also redesign clinical workflows, obtain regulatory clearance, retrain

[7] "Production" here refers specifically to the use of AI in producing goods or services, as defined by the US Census Bureau's Business Trends and Outlook Survey, which covers nonfarm employer businesses across all sectors and firm sizes. It excludes AI use in other business functions such as marketing, logistics, or administration. This narrow definition accounts for much of the apparent disconnect between worker-level and firm-level adoption rates: when the Census broadened the question in November 2025 to cover any business function, the share of US firms using AI rose to 17 percent, and Bick et al. (2026) estimate that on a basis comparable to European firm surveys it would be roughly 34 percent. Two further factors contribute to the remaining gap. The firm-level rate is an unweighted share of firms and thus dominated by small businesses, whereas AI-using workers are concentrated in large firms. In addition, some workers use AI at work without formal adoption by their employer.

staff, and address liability structures. A law firm would have to develop quality-control procedures and navigate evolving professional-responsibility guidance.

Moreover, the productivity gains from other general-purpose technologies have accelerated once complementary products emerged that made reorganization practical: standardized electric motors and modular machine tools for electrification; enterprise software and networking equipment for information technology. In the artificial intelligence market, an analogous ecosystem is developing at a much faster pace. Increasingly, intelligence is bundled into software systems that perform complex, multi-step tasks on behalf of users. For example, tools such as Claude Code and Codex bundle model inference with execution logic and workflow integration, allowing users to delegate complex tasks rather than issue individual prompts. By reducing integration costs, these innovations expand the range of economically viable uses.

While the market described in this paper has developed the models and inference technologies and made them widely accessible, the complementary applications that translate these capabilities into usable products remain in their infancy, except for a few industries. Currently, a large ecosystem of startups—backed by billions of dollars in venture capital—is rapidly developing complementary applications. How quickly these markets produce compelling products—and how rapidly firms adapt their internal processes to integrate them—will be central to determining AI's ultimate economic impact.

Where Is the Market Headed?

The market for machine intelligence is only a few years old, yet it has already developed a layered supply chain, attracted dozens of competing model creators and nearly a hundred inference providers, and driven the quality-adjusted price of intelligence down by three orders of magnitude. How both the technology and the market evolve from here is a multi-trillion-dollar question and one that calls for humility.

A central question is whether and how soon the AI market may consolidate. Early signs of stabilization have already appeared: the set of frontier model creators has remained largely unchanged in the latter part of 2025 and into early 2026, even as innovation continues. If training costs keep rising, fewer firms will be able to compete at the frontier, given that state-of-the-art runs now cost hundreds of millions of dollars (Emberson and Edelman 2025). But these barriers to entry may not last. Open-source models are close behind, and improvements in algorithms may reduce training and deployment costs. There is also a possibility that local devices, such as phones, will be able to serve at least some inference needs without relying on cloud providers. In any of these cases, costs may collapse even further.

For most of economic history, intelligence was bundled with human labor and allocated through labor markets. Today, machine intelligence is being unbundled from humans, supplied through its own markets, and purchased on demand. Most

informed observers expect frontier capabilities to keep improving while prices continue to fall (Murphy et al. 2025). As intelligence becomes cheaper and more capable, it will become a standard input into production, and more tasks will be automated. The declining cost of intelligence may therefore alter the boundaries of firms and the structure of labor demand. It remains to be seen whether these changes quickly translate to productivity or whether bottlenecks slow this process down.

■ *We thank Alessandro Bonatti, Seth Benzell, James Brand, Erik Brynjolfsson, Tom Cunningham, Joseph Doyle, Dean Eckles, Garrett Johnson, Daniel Rock, Daniel Waldinger, the JEP editors, and seminar participants at the Stanford Digital Economy Lab, MIT Sloan, and MIT FutureTech. We thank Marilyn Pareboom for exceptional research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect the views of Amazon or Microsoft. None of the authors hold a financial interest in Artificial Analysis or OpenRouter. To our knowledge, Amazon and Microsoft have no material financial interest in Artificial Analysis or OpenRouter.*

References

Acemoglu, Daron. 2025. "The Simple Macroeconomics of AI." *Economic Policy* 40 (121): 13–58.

Acemoglu, Daron, and Pascual Restrepo. 2019. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." *Journal of Economic Perspectives* 33 (2): 3–30.

AI Impacts. 2015. "Trends in the Cost of Computing." https://aiimpacts.org/trends-in-the-cost-of-computing/#Wikipedia.

AI Impacts. 2017. "Wikipedia history of GFLOPS costs." https://aiimpacts.org/wikipedia-history-of-gflops-costs/.

Aubakirova, Malika, Alex Atallah, Chris Clark, Justin Summerville, and Anjney Midha. 2025. *State of AI: An Empirical 100 Trillion Token Study with OpenRouter.* OpenRouter.

Bick, Alexander, Adam Blandin, and David J. Deming. 2024. "The Rapid Adoption of Generative AI." NBER Working Paper 32966.

Bick, Alexander, Adam Blandin, David J. Deming, Nicola Fuchs-Schündeln, and Jonas Jessen. 2026. "Mind the Gap: AI Adoption in Europe and the U.S." NBER Working Paper 34995.

Bresnahan, Timothy F., and M. Trajtenberg. 1995. "General Purpose Technologies: 'Engines of Growth'?" *Journal of Econometrics* 65 (1): 83–108.

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

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. "Generative AI at Work." *Quarterly Journal of Economics* 140 (2): 889–942.

Chatterji, Aaron, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. "How People Use ChatGPT." NBER Working Paper 34255.

Comin, Diego, and Bart Hobijn. 2010. "An Exploration of Technology Diffusion." *American Economic Review* 100 (5): 2031–59.

David, Paul A. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *American Economic Review* 80 (2): 355–61.

Demirer, Mert, Andrey Fradkin, and Nadav Tadelis. 2026. *Data and Code for: "The Emerging Market for Intelligence: How Firms Buy and Sell AI."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/EX249283V1.

Demirer, Mert, Andrey Fradkin, Nadav Tadelis, and Sida Peng. 2025. "The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs." NBER Working Paper 34608.

DrPeering. 2016. "Internet Transit Price Declines." https://drpeering.net/HTML_IPP/chapters/ch02-Internet-Transit/ch02.1-Internet-Transit-Prices.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs Are GPTs: Labor Market Impact Potential of LLMs." *Science* 384 (6702): 1306–08.

Emberson, Luke, and Yafah Edelman. 2025. "Compute accounts for the majority of expenses of AI companies." Epoch AI. https://epoch.ai/data-insights/company-spending-breakdown.

Greenstein, Shane. 2015. *How the Internet Became Commercial: Innovation, Privatization, and the Birth of a New Network*. Princeton University Press.

Handa, Kunal, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, et al. 2025. "Which Economic Tasks Are Performed with AI? Evidence from Millions of Claude Conversations." Preprint, arXiv. https://doi.org/10.48550/arXiv.2503.04761.

Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. "Why Language Models Hallucinate." Preprint, arXiv. https://doi.org/10.48550/arXiv.2509.04664.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." Preprint, arXiv. https://doi.org/10.48550/arXiv.2001.08361.

Klepper, Steven. 1996. "Entry, Exit, Growth, and Innovation over the Product Life Cycle." *American Economic Review* 86 (3): 562–83.

Klepper, Steven. 2002. "Firm Survival and the Evolution of Oligopoly." *RAND Journal of Economics* 33 (1): 37–61.

Lerner, Josh, and Jean Tirole. 2005. "The Economics of Technology Sharing: Open Source and Beyond." *Journal of Economic Perspectives* 19 (2): 99–120.

Mathieu, Edouard. 2023. "The price of computer storage has fallen exponentially since the 1950s." Our World in Data.

Moore, Gordon E. 1965. "Cramming More Components onto Integrated Circuits." *Electronics Magazine* 38 (8): 114–17.

Murphy, Connacher, Josh Rosenberg, Jordan Canedy, Zach Jacobs, Nadja Flechner, Rhiannon Britt, Alexa Pan, et al. 2025. "The Longitudinal Expert AI Panel: Understanding Expert Views on AI Capabilities, Adoption, and Impact." Forecasting Research Institute Working Paper 5.

Nagle, Frank, and Daniel Yue. 2025. "The Latent Role of Open Models in the AI Economy." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.5767103.

National Human Genome Research Institute. 2022. "The Cost of Sequencing a Human Genome." https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

Papailiopoulos, Dimitris. 2026. "You Don't Need to Run Every Eval." February 25. https://dimitrisp.substack.com/p/you-dont-need-to-run-every-eval.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot." Preprint, arXiv. https://doi.org/10.48550/arXiv.2302.06590.

Ritchie, Hannah, and Pablo Rosado. 2026. "Battery costs have declined by 99% in the last three decades, making electrified transport a reality." Our World in Data. https://archive.ourworldindata.org/20260330-065433/battery-price-decline.html

Ritchie, Hannah, Pablo Rosado, and Max Roser. 2023. "Solar photovoltaic module price." Our World in Data. https://archive.ourworldindata.org/20260518-093348/grapher/solar-pv-prices.html.

Roser, Max, Hannah Ritchie, and Edouard Mathieu. 2023. *Data Page: "Cost of Sequencing a Full Human Genome."* Data adapted from National Human Genome Research Institute. https://ourworldindata.org/grapher/cost-of-sequencing-a-full-human-genome.

Sarkar, Suproteem K. 2025. "AI Agents, Productivity, and Higher-Order Thinking: Early Evidence from Software Development." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.5713646.

Tully, Tim, Joff Redfern, Deedy Das, and Derek Xiao. 2025. *2025 Mid-Year LLM Market Update: Foundation Model Landscape + Economics*. Menlo Ventures.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 47–72

US Tariff Policy Since 1789

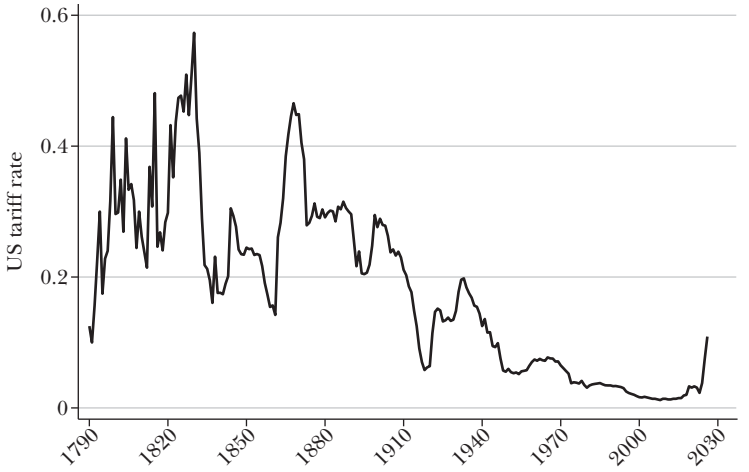
Miguel Acosta, Lydia Cox, Andrew Greenland, John Lopresti, Christopher M. Meissner, Martin Rotemberg, and Sharon Traiberman

Despite the centrality of trade policy in the economic history of the United States, the data and historical perspective required to explore its effects in detail have, until now, been in too short supply. While a substantial literature has focused on rigorously quantifying the effects of tariff liberalization (Romalis 2007; Caliendo and Parro 2015; Hakobyan and McLaren 2016; Pierce and Schott 2016; Amiti, Redding, and Weinstein 2019; Fajgelbaum, Goldberg, Kennedy, and Khandelwal 2020), researchers have largely focused on trade agreements in the 1990s and later—a period in which both the structure and magnitude of tariff changes differ substantially from those that characterize the majority of US history. This is true even of studies focused on the tariff increases of 2025, which yielded

■ *Miguel Acosta is Assistant Professor of Economics, University of Wisconsin, Madison, Wisconsin. During the 2025–2026 academic year, he was a Visiting Assistant Professor of Economics, Princeton University, Princeton, New Jersey. Lydia Cox is Assistant Professor of Economics, University of Wisconsin, Madison, Wisconsin. During the 2025–2026 academic year, she was an IES Kenen Fellow, Princeton University, Princeton, New Jersey. Andrew Greenland is Assistant Professor of Economics, Department of Agriculture and Resource Economics, College of Agriculture and Life Sciences, and Department of Economics, Poole College of Management, North Carolina State University, Raleigh, North Carolina. John Lopresti is Associate Professor of Economics, William & Mary, Williamsburg, Virginia. Christopher M. Meissner is Professor of Economics, University of California–Davis, Davis, California. Martin Rotemberg is Associate Professor of Economics, New York University, New York City, New York. Sharon Traiberman is Assistant Professor of Economics, New York University, New York City, New York. Greenland is the corresponding author at agreenl@ncsu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20251488.

Figure 1
US Tariff Ad Valorem Equivalent Tariff Rate, Since 1790



Source: Various data sources including the Historical Statistics of the United States, USITC, and for data after 2016 the Trade War Tracker.
Note: Figure provides annual US import-weighted ad valorem equivalent tariff rate from 1790–2026.

tariff rates that, while high by modern standards, remained substantially lower than those observed throughout most of US history, as displayed in Figure 1. On the other hand, authoritative work by Taussig (1910) and Irwin (2017), among others, provides sweeping perspective on the evolution and determinants of trade policy over time, but has had to do so without access to comprehensive micro-data. This article seeks to serve as a bridge between these two strands of literature. We rely on a newly assembled, highly disaggregated historical tariff dataset—the construction of which has been generously funded by the National Science Foundation—to introduce readers to several underappreciated aspects of the history of US trade policy. We use these data to shed light on how tariffs affect the American economy and how trade policy today contrasts with historical precedents.

We begin with a historical overview of the three "Rs" of historical tariff policy as proposed by Irwin (2017): the use of tariffs as a key revenue stream prior to the Civil War; the restrictive tariffs of the newly emergent industrial economy after the Civil War; and the period of "reciprocity" that emerged beginning in the 1930s. After this overview, we use our data to provide insights into the mechanics of tariff policy over the long run. In particular, we emphasize that the state capacity to administer tariffs has evolved over time, which has enhanced US capacity both to strike new trade deals and to satisfy political-economy and distributional concerns associated with such laws. While tariff rates have broadly decreased over time and are now substantially lower than they were 150 years ago, we also show substantial variation across sectors. Tariff rates for food and crude materials have fallen fairly consistently

since the Civil War. Meanwhile, apparel and machinery occasionally experienced substantial increases in their rates, especially in the 1890s and 1930s.

We then call attention to the use of specific tariffs and the implications of this well-known but underanalyzed feature of US trade policy. Specific tariffs—defined in terms of physical units of a product—while still in use today, were once the principal form in which tariffs were set. Unlike ad valorem tariffs, specific tariffs have distinct implications for revenue collected and for the impact of trade policy on American industry and consumers. Importantly, the ad valorem equivalent of a specific tariff depends on the price level, so the protective effect of US trade policy has often moved substantially and inversely with price levels. Using a simple decomposition, we show that, because specific tariffs were ubiquitous in US tariff laws, aggregate price changes mechanically moved aggregate tariff rates and are associated with a large share of cross-decade tariff variation. With this in mind, we discuss ongoing work rationalizing the use of and declining reliance on specific tariffs and what their price-level sensitivity implies for Congress's attempts to protect particular industries and for researchers exploring the effects of tariff policy on micro- and macroeconomic outcomes.

In the penultimate section, we introduce our new dataset, representing the first comprehensive collection of US historical tariff lines, imports, and their reliance on specific and ad valorem tariffs.[1] This dataset constitutes the universe of US tariff policy since the country's inception. We show how new research based on these data and the insights discussed earlier can shed light on the impact of tariffs over the long run and can be used to address endogeneity problems that plague most attempts to assess the impact of trade policy. We illustrate the importance of having access to comprehensive and granular data for our understanding of how tariffs affected late nineteenth century industrialization, early twentieth century structural change, and the impact of trade policy on international competition and inequality in the mid-to-late twentieth century. These historical studies suggest that tariffs can have important implications for the American economy in some ways that have long been appreciated and in other ways that are more novel.

The Shifting Institutions of Tariff Protection

Debates over tariffs and their distribution across industries date back to the earliest days of American history. As Irwin (2017) writes, "within just a few days of the opening of Congress, the great debate over trade policy was joined. That perennial debate revolves around the proper objective of import duties: to raise

[1] While the analysis in the present article is based on snapshots of our database that have been aggregated over sectors and time, the ultimate disaggregate annual trade and tariff database will be released at the end of the grant period. Individual components may become available sooner as the primary studies from which the data were produced and their replication packages become available (https://www.nsf.gov/awardsearch/show-award/?AWD_ID=2314696&HistoricalAwards=false).

revenue, to restrict imports in order to protect domestic manufacturers, or to achieve reciprocity—or some combination of the three." Frankly, this debate among policymakers has evolved very little since the nation's founding. However, the institutions under which US trade policy is determined have evolved.

For the first 150 years of US history, US tariffs were determined by the US Congress. Their levels and distribution across industries were determined in response to changing economic needs and political preferences. Beginning in the 1930s, this pattern shifted, with US tariffs largely determined inside of a multilateral system in which the ideas of reciprocity and nondiscrimination were central. In the post–World War II period, and eventually under the World Trade Organization, reciprocity involved reductions in tariffs and other concessions that led to an increase in imports comparable to the rise in exports achieved due to tariff reductions by trade partners. During President Trump's first term, and more forcefully in his second term, he began a legally contested process of changing tariffs according to his own notions of reciprocity. This latest version of reciprocity is somewhat reminiscent in spirit to the idea championed in the late nineteenth century by President William McKinley, which allowed for the possibility of lower US tariffs provided that trade partners did not discriminate against US imports.

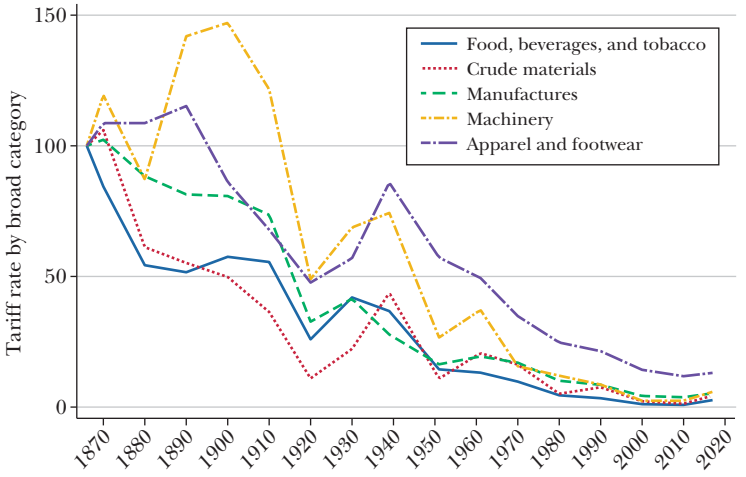
Tariffs for Revenue

Prior to the US Civil War, tariffs were the fundamental tool of the US government for raising revenue. To accomplish this goal, tariffs were imposed on products like alcohols, tea, coffee, and sugar—many of which were not produced in the United States but were imported. The need for revenue was especially acute when the US government needed to pay wartime debts, and so tariffs were generally higher for stretches after the Revolutionary War, the War of 1812, and the Civil War. The normal objective was for a peacetime balanced budget, which implied that tariff rates were likely to respond counter to the trajectory of the budget (Taussig 1910). Over the course of the pre–Civil War era, tariffs accounted for approximately 90 percent of government revenue and 2 percent of GDP (United States Department of the Treasury 1980; Johnston and Williamson 2026).

The Rise of the Protective Tariff

In addition to raising revenue, nineteenth-century political debates on tariff policy centered around the extent to which tariffs should also be used to support domestic manufacturing production. Representatives from agrarian states—mostly concentrated in the South and later the Midwest—tended to support lower import duties. Because the United States was primarily an exporter of agricultural goods, import duties raised prices of consumption goods for residents of those regions. Also, if other countries retaliated, US tariffs threatened the ability of farmers to access foreign markets for their goods. On the other hand, manufacturing regions—primarily the Mid-Atlantic and later the Northeast—felt the threat of foreign competition and wanted as much tariff protection as possible. Proponents of reducing duties argued that a system of protecting industries would surely lead

Figure 2
Tariff Rate for Broad Categories



Source: Acosta et al. (2026a).
Note: This figure shows the ad valorem equivalent tariff rate for broad industry groups. Food/Beverages/Tobacco is made up of Sections 0 (food and live animals) and 1 (beverages and tobacco) of the SITC revision 2. Crude materials contains SITC Sections 2 (crude materials, inedible), 3 (mineral fuels, lubricants, and related materials), 4 (animal and vegetable oils, fats, and waxes), and 5 (chemicals and related products). Manufactures represents SITC Section 6 (manufactured goods classified chiefly by material), Machinery Section 7 (machinery and transport equipment), and Apparel/Footwear Section 8 (miscellaneous manufactured articles).

to corruption, lining the pockets of those that were able to influence the government, while proponents of higher tariffs argued that foreign competition would be detrimental for import-competing industries (Irwin 2017).

Because the authority to tax and tariff had been granted to the legislative branch in the US Constitution, debates on the appropriate level of tariffs across industries occurred in Congress, and the tariffs enacted were largely based on which party was in power. The ebb and flow of these interests and budgetary concerns are apparent in the general patterns in Figure 2, which displays ad valorem equivalent tariffs, indexed to 1865 levels, for a few broad industry categories. For example, between 1865 and 1890, when the Republican (protectionist) party was in control, tariffs on manufactured goods, machinery, apparel, and footwear remained elevated. On the other hand, tariffs on many of the traditional revenue products, like coffee, tea, tobacco, and sugar (contained in the Food/Beverages/Tobacco category), were reduced in an effort to reduce the government surplus generated by tariff revenue.

Republicans remained in control of Congress and the executive branch in most years of the late nineteenth century and therefore maintained their high-tariff regime until Woodrow Wilson took office in 1913, with the notable exception being the term of President Grover Cleveland from 1892 to 1896. In 1894, a significant

reduction in tariffs occurred with the so-called Wilson-Gorman tariff. Later under President Wilson, the Underwood-Simmons tariff was passed in 1913, representing a Democrat-led broad-based reduction in tariffs. Under the new legislation, across-the-board reductions of around 25 percent were applied to manufactured goods, and many agricultural products, raw materials, and manufactured goods were moved to the duty-free list. This decline in legislated rates in 1913 generated a sharp decline in ad valorem equivalent rates, as can be seen in Figure 2.[2] The income tax, newly created in 1913, helped to offset the revenue losses to the federal government.

Bilateralism, Multilateralism, and Tariff Rate Heterogeneity

A fundamental change to how the US government set tariffs occurred with the passage of the Reciprocal Trade Agreements Act of 1934. This institutional reform gave the executive the authority to negotiate with foreign nations to reduce tariffs in return for reciprocal reductions in tariffs in the United States. The passage of the Reciprocal Trade Agreements Act was an acknowledgment by the administration of President Franklin Roosevelt that if the United States wanted to be able to sell its goods abroad—considered crucial for economic recovery from the Great Depression—it must also be willing to buy goods from other countries (Roosevelt 1934). As a result, the US tariff schedule was no longer solely a reflection of domestic interests and party politics, but instead began to reflect the interests of US trading partners as well. While the fragile state of the US economy coming out of the Great Depression prevented large broad-based reductions in tariffs during this period, Acosta and Cox (2025) show that bilateral negotiations generated increased variation in tariff rates within narrow categories of goods. The United States reduced tariffs, but restricted those reductions to the varieties in which its trading partners were interested. We will return to this increasing granularity of tariffs in the next section.

Bilateralism turned into multilateralism beginning in 1947 with the General Agreement on Tariffs and Trade (GATT)—an agreement between nations to reduce trade barriers implemented through rounds of negotiations. A large body of work has studied these multilateral negotiations from a theoretical perspective (for example, Bagwell and Staiger 1999, 2002; Maggi 1999; Ossa 2011; Bagwell, Staiger, and Yurukoglu 2020). However, our disaggregated data provide a new window into how the nature of these negotiations determined the distribution of tariff rates between industries. Between 1947 and 1963, the GATT bargaining rounds shared a continued focus on reducing global tariff rates through line-by-line negotiations, much like the bilateral negotiations during the period immediately following the Reciprocal Trade Agreements Act of 1934.

One important feature of this approach is that, because the tariffs were determined through reciprocal agreements, they became much harder to adjust. Starting with the Kennedy Round of the GATT, which began in 1964, countries agreed to

[2] However, we note that not all of the decline in this figure, or of aggregate ad valorem equivalent tariff rates in general, reflect legislated tariff changes—we return to this point shortly.

make broad-based reductions of their entire tariff schedules (subject to exceptions), instead of engaging in line-by-line tariff negotiations. This change in format meant that preexisting cross-sectional variation in tariffs was largely preserved, even as average tariff levels came down. This pattern can also be seen in Figure 2: beginning in the 1960s, there is a general convergence of rates toward zero. Indeed, starting with the Tokyo Round of the GATT in 1973, reductions were made according to the "Swiss formula"—a nonlinear formula in which high tariff rates were cut more aggressively than low tariff rates. With these changes, variation in tariff levels across goods declined.

An awareness of the institutions that determine trade policy is important to understanding both the patterns and effects of protection. A high tariff rate may reflect tariffs imposed on products for revenue purposes, to appease lobbyists, for industrial policy motives, to protect inputs or final goods, or simple path dependence. The disaggregated tariff database that we have constructed can help distinguish these various channels, providing a new window through which we can more accurately understand the effects of trade policy.

Increasing Granularity of Protection

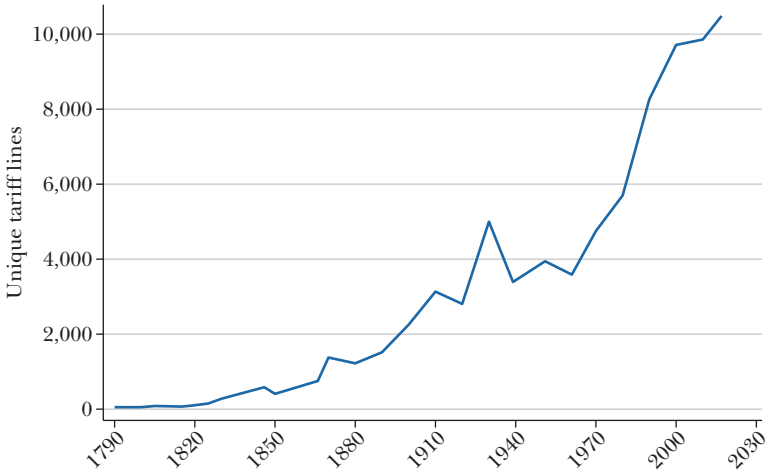
Tariffs are imposed at what is commonly referred to as the "tariff-line" level. Each tariff line corresponds to either an individual product or a group of nondelineated but similar products. For example, at present there are two tariff lines associated with hewing tools: one for machetes (with no tariff) and one for other hewing tools (like axes, with a 4.5 percent tariff). Shifts in the average US tariff mask the exponential growth in the granularity of US tariffs over time. Figure 3 shows the number of distinct tariff lines in the United States between 1790 and today. As shown in the figure, the number of tariff lines has increased from fewer than 100 in 1790 to over 10,000 today. In the early years, this vast expansion reflected an increase in the administrative capacity of the US government. More recently, it reflects an effort by policymakers to increase the precision with which changes in tariff policy are applied.

1789–1850: Broad Tariffs, Inconsistent Reporting

In the very first session of the US Congress, the Tariff Act of July 4, 1789, imposed duties on imports in order to generate federal revenue (Irwin 2017). Within the same month, Congress passed a comprehensive collection statute that constructed a port bureaucracy—collectors, naval officers, surveyors, and inspectors—and required manifests, entries, and sworn documents to establish the tax base. A decade later, Congress consolidated and expanded these procedures in the Collection Act of 1799, which became the backbone of customs administration (Goss 1897; Madsen, Rotemberg, Traiberman, and Wang 2025).

During these early years, legislated tariff bills were a far cry from the organized schedules of detailed product-level tariff rates that we have today. The use of

Figure 3
Number of Tariff Lines, 1790–2017



Source: Acosta et al. (2026a).
Note: This figure shows the number of distinct tariff lines in years between 1790 and 2017. Between 1790 and 1850, the number corresponds to the number of distinctly reported imports in US import statistics. In the following years, only tariff lines on products with positive imports are included in the figure.

industry codes to organize the tariff schedule was not considered before the US Civil War. Instead, tariff bills were written in plain text, without a one-to-one mapping to goods. For example, in some years all woolen manufactures received the same tariff, and in other years, rugs and clothing received different tariffs. Customs officials would occasionally compile the set of current tariffs and sell them to other officials or business (Jones 1835, 1854), but contemporaneous accounts emphasize the "great labor" required to do so (Madsen, Rotemberg, Traiberman, and Wang 2025). In these years, the information on tariffs collected and published represented a bare minimum. In tariff bills, goods were typically not named at all; instead, Treasury reported total imports by tariff rate category (for example, total value paying a 5 percent duty). Other information irrelevant to tariff assessment was generally not collected: quantities were not recorded for goods subject to ad valorem tariffs, values were not recorded for goods subject to specific duties, and goods on the no-tariff "free list" were often omitted entirely.

A major institutional change for data collection on tariffs occurred with the Act of February 10, 1820, as it was named, which directed the Treasury to produce annual statements showing the "actual state of commerce and navigation" between the United States and foreign countries in forms prescribed by the Secretary of the Treasury (US Congress 1820). These statements were published as *Commerce and Navigation*. This change led to a partial improvement in the quality of data collection and good-level reporting in import statistics. Some quantities began to be reported for ad valorem goods, values were increasingly reported for

specific-duty goods, and free-list items appeared more regularly. These changes are reflected in Figure 3, where in the period between 1820 and 1850 there is an almost quadrupling of the number of distinctly reported goods in import statistics. These improvements in reported trade data were not costless, requiring clerical labor, bookkeeping, and standardized routines at the ports (Jones 1835). Indeed, the share of customs personnel titled clerk or bookkeeper was near-zero in 1816, but had reached 15 percent by early in the 1820s *Official Register of the United States* (US Department of State 1816–1850). By 1850, clerks and bookkeepers accounted for roughly one-fifth of customs employment. Note that this increase in share was accompanied by increases in overall Customs employment, which rose from around 190 employees in 1817 to almost 2000 in 1850.

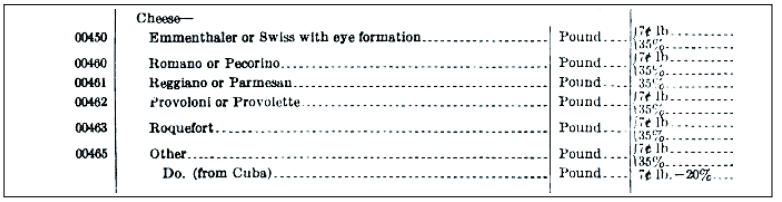
From 1850 to the Modern Era: More Tariff Lines, More Policy Precision

The number of tariff lines continued to steadily increase throughout the nineteenth and twentieth centuries. As shown in Figure 3, the number of tariff lines roughly quadrupled again between 1850 and 1930, before increasing another five-fold between 1930 and today. Grant (2023) proposes a model of endogenous classification that broadly fits the patterns we see in the data. In his model, policymakers choose the level of disaggregation at which to enforce policies like tariffs by trading off the costs of enforcing a more complex policy—including costs of gathering information—with the perceived political benefits of more targeted policy.

That policymakers saw a benefit in more targeted tariffs can be seen most clearly from the early 1930s to today. Early in this period, which coincided with the rise of US involvement in first bilateral and then multilateral trade negotiations, policymakers divided tariff lines into more detailed categories in order to target more precisely products relevant to the countries with which the United States was negotiating. For example, Figure 4 shows that in 1930, imports of cheese were broken into six categories, all of which faced a tariff rate of 7 cents per pound with a minimum of 35 percent. By 1937, having negotiated agreements with Switzerland, the Netherlands, Finland, Canada, and France, the number of cheese-related tariff lines in the tariff schedule—shown in Figure 5—had doubled. In the agreement with Switzerland, a new tariff line for "Gruyere process-cheese" was created and the tariff rate was reduced to 7 cents per pound with a 20 percent minimum for that specific variety. In the agreement with Canada, new tariff lines for "Cheddar" were created and their tariff rates lowered from the original level, and similarly for the agreements with France and the creation of a tariff line for "blue-mold" cheese and with the Netherlands and the creation of a line for "Edam and Gouda cheeses."

By creating additional lines, the United States was able to lower rates on some types of cheese while leaving the tariff on "other" types of cheese intact. In fact, in describing the agreement negotiated with the United Kingdom in 1939, the US Tariff Commission stated in its *Annual Report* that "many new import classifications are established by the agreement either for the purpose of confining concessions to products supplied principally by the United Kingdom, or in order

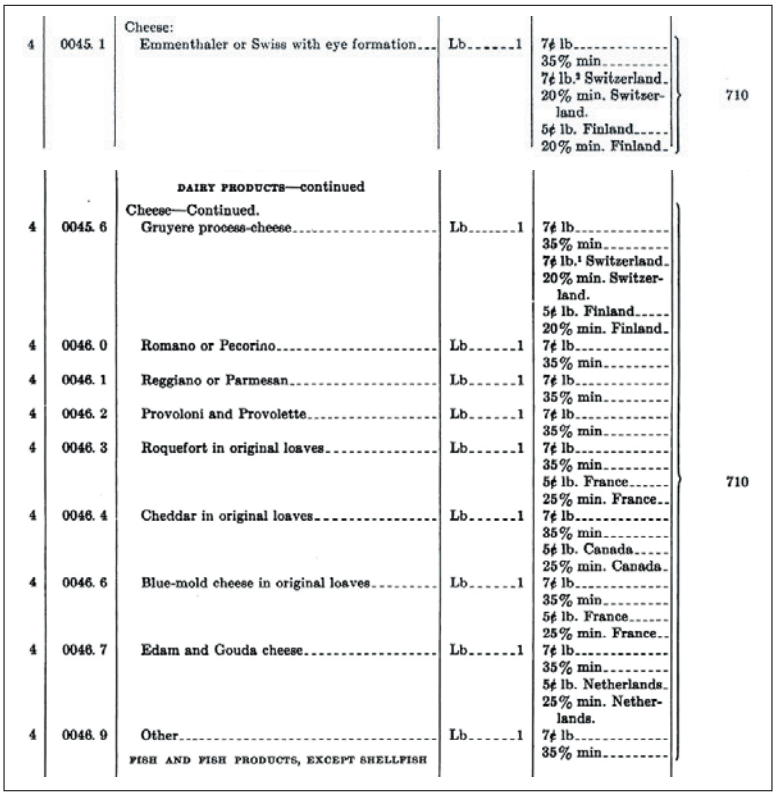
Figure 4
1930 Schedule A: Cheese



	Cheese—		
00450	Emmenthaler or Swiss with eye formation	Pound	7¢ lb 35%
00460	Romano or Pecorino	Pound	7¢ lb 35%
00461	Reggiano or Parmesan	Pound	35%
00462	Provoloni or Provolette	Pound	7¢ lb 35%
00463	Roquefort	Pound	7¢ lb 35%
00465	Other	Pound	7¢ lb 35%
	Do. (from Cuba)	Pound	7¢ lb.—20%

Source: Schedule A. Statistical Classification of Imports into the United States. US Department of Commerce (1930).
Note: This figure shows the portion of the 1930 tariff schedule covering varieties of cheese.

Figure 5
1937 Schedule A: Cheese



4	0045. 1	Cheese: Emmenthaler or Swiss with eye formation	Lb.—1	7¢ lb 35% min 7¢ lb.[3] Switzerland 20% min. Switzerland. 5¢ lb. Finland 20% min. Finland	710
		DAIRY PRODUCTS—continued			
		Cheese—Continued.			
4	0045. 6	Gruyere process-cheese	Lb.—1	7¢ lb 35% min 7¢ lb.[1] Switzerland 20% min. Switzerland. 5¢ lb. Finland 20% min. Finland	
4	0046. 0	Romano or Pecorino	Lb.—1	7¢ lb 35% min	
4	0046. 1	Reggiano or Parmesan	Lb.—1	7¢ lb 35% min	
4	0046. 2	Provoloni and Provolette	Lb.—1	7¢ lb 35% min	
4	0046. 3	Roquefort in original loaves	Lb.—1	7¢ lb 35% min 5¢ lb. France 25% min. France	710
4	0046. 4	Cheddar in original loaves	Lb.—1	7¢ lb 35% min 5¢ lb. Canada 25% min. Canada	
4	0046. 6	Blue-mold cheese in original loaves	Lb.—1	7¢ lb 35% min 5¢ lb. France 25% min. France	
4	0046. 7	Edam and Gouda cheese	Lb.—1	7¢ lb 35% min 5¢ lb. Netherlands 25% min. Netherlands.	
4	0046. 9	Other	Lb.—1	7¢ lb 35% min	
		FISH AND FISH PRODUCTS, EXCEPT SHELLFISH			

Source: Schedule A. Statistical Classification of Imports into the United States. US Department of Commerce (1937).
Note: This figure shows the portion of the 1937 tariff schedule covering varieties of cheese. While reduced rates are noted along with the country with which those reduced rates were negotiated, the "most favored nation" (MFN) principle applied and so reductions were extended to all imports.

to limit duty reductions to the less competitive portions of the old classifications" (US Tariff Commission 1939). This type of behavior was systematic, especially during the period between 1930 and the mid-1960s (Acosta and Cox 2025).

The second part of the above quotation alludes to new tariff lines that were created as a way to improve the targeting of tariff policy across products, not just across countries. This behavior becomes particularly apparent as new products are invented or as recent inventions become mainstream. For example, prior to 1960, there was no separate tariff line for computers—they were instead included in a catch-all category covering "Other machines having as an essential feature an electrical element or device." In 1960, however, a new tariff line was created covering "Electronic computers and related electrical or electronic information processing machines, and parts thereof." New tariff lines continue to be created for new products or products that have become increasingly important. In the most recent revision of the Harmonized System (HS2022), an internationally standardized set of codes used for classifying traded products, the United States now has tariff lines for electronic waste, drones, and smartphones.[3] Tariffs began as a broad-based tax designed to raise revenue, but have since evolved into a system for tracking highly detailed products and enforcing policy in a targeted way across countries and products.

Complexity, Legislated Tariffs, and Enforced Tariffs

The interaction between the complexity of the tariff schedule and the administrative capacity of US customs determines the extent to which legislated tariff policy is actually enforced. From the first Congress through much of the early 1800s, despite having developed a fiscal system that depended on customs revenue, the US government could not reliably reconcile what was owed with what had been collected. In April 1816, Representative Benjamin Huger chaired a House of Representatives inquiry into "unsettled balances" due to the United States (Rao 2016). Overlapping statutes, inconsistent Treasury circulars, and idiosyncratic port practices created what Huger described as a "labyrinth," with problems "especially among the collectors." Huger concluded that Congress could not exercise meaningful oversight without more standardized reporting. The Committee's proposed remedy was "annual and detailed reports" that would provide the Treasury with better practical information (Huger 1816). Huger also emphasized that it was "impossible to anticipate all the difficulties" that would arise with the evolution of Customs over time.

The lack of reliable records directly reduced duties collected, particularly through the administration of "duty bonds," under which importers promised to pay duties after selling their wares. In 1807, Baltimore's collector reviewed customs bond records and found that previous collectors had not even maintained a list of

[3]Prior to the Harmonized System (HS), classification was not standardized across countries. HS codes are classified and subclassified down to a six-digit level by the World Customs Organization (WCO)—so the introduction of new product codes for e-waste, drones, and smartphones was actually done at the six-digit level by the WCO. Countries can then add additional digits that are not standardized for legislation or information collection. The United States legislates tariffs at the eight-digit level, but collects import statistics up to ten digits.

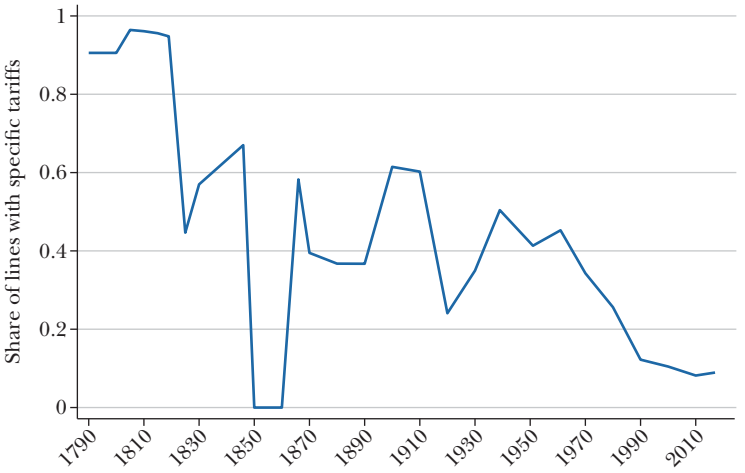
overdue bonds. Conditions were worse at Mobile: by 1811, its collector admitted that not one duty bond had been collected in the District since its establishment (Rao 2016). Weak recordkeeping also facilitated systematic undervaluation. In the 1820s, the dollar value of exported spices on which drawbacks—rebates on duties for imported goods that are then exported elsewhere—were claimed exceeded the dollar value of spice imports, despite the fact that spices were not produced domestically and were extensively consumed in the United States (Goss 1897). Such an outcome would have been impossible had export shipments been reliably linked to the corresponding import records that generated drawback eligibility.

The most dramatic failure of Customs accounting occurred in the defalcation of Samuel Swartwout, Collector of Customs for the port of New York. During his tenure, Swartwout systematically withheld funds that should have been remitted to the US Treasury, including tonnage duties, portions of fines and forfeitures, and collections on bonds. He reported large sums "on hand," justifying their retention as necessary to cover office expenses, pending drawbacks, protested duties, and unsettled accounts, while failing to provide a verifiable reconciliation. When his term ended in 1838 and he departed for Europe, Treasury investigations estimated the resulting shortfall at over $1.3 million in nominal 1838 dollars—at a time when federal receipts in the 1830s ranged from $25 to $35 million (Congressional Globe 1838).

Congress demanded to know why Treasury could not readily ascertain the defalcation "when it first commenced." In a January 1839 response, the Secretary of the Treasury offered two explanations (US Treasury Department 1839). First, the information flow that might have revealed the problem was not a statutory audit system: Weekly and monthly returns from collectors were described as voluntary instructions "not required by any law," maintained for administrative convenience rather than as a formal control mechanism. Second, even this voluntary paper trail was institutionally fragmented and operationally unwieldy. Treasury emphasized that the relevant New York returns were so numerous and dispersed across offices that reproducing them would have absorbed clerical capacity for months, crowding out current operations.

Nearly two centuries later, with entirely different policies and systems in place, the United States is again being confronted with the fact that tariff policy is only enforceable to the extent that the administrative apparatus is able to keep track of key information—what arrived, who imported it, how it is valued, what is owed, and what has been paid. It has been reported that the volume and complexity of changes to tariff rates imposed by the Trump administration starting in 2025 has made it difficult for the agencies responsible for screening imports and collecting duties to enforce those changes, making it easier for importers to evade the charges (Feng 2025). In this symposium, Gopinath and Neiman suggest that this heterogeneous enforcement has contributed to a gap between legislated and actual rates, moderating the impact that the increase in tariffs has had on the economy. Whether, as proposed by Krueger (1974), this kind of selective enforcement can lead to additional economic costs as firms lobby politicians to protect or create rents, has to our knowledge, gone largely unexplored.

Figure 6
Specific Tariff Share of Tariff Lines, 1790–2017



Source: Acosta et al. (2026a).
Note: This figure shows the share of distinct tariff lines that have either specific or compound tariff rates in years between 1790 and 2017. Between 1790 and 1850, the number corresponds to the number of distinctly reported imports covered by specific tariffs in US import statistics. In the following years, only tariff lines on products with positive imports are included in the figure.

The Instruments of Protection: Specific and Ad Valorem Tariffs

Thus far, we have not drawn attention to what type of tariffs policymakers employed, and readers may presume that all tariffs are charged as a percent of the value of goods imported. This perception and treatment spill over into the academic literature and popular press as well. Indeed, the most common representation of tariffs—and all of our prior figures—reports the aggregate tariff rates on a percent basis. Yet this convention is a matter of convenience, and it masks interesting and important shifts in policy. How much does the type of protection matter for the US economy and for how one should think about the primary determinants of protection and liberalization? We turn to this question now.

Tariffs come in two primary forms: specific tariffs are imposed as a price per unit—for example, dollars per pound—while ad valorem tariffs are imposed as a percentage of value. Compound tariffs are a combination of specific and ad valorem rates—say, 10 percent plus 5 cents per pound. In modern research on tariff policy, it is conventional to treat tariffs as ad valorem, but for much of US history, specific and compound tariffs were the dominant forms of protection (Irwin 1998, 2017). Figure 6, which shows the percent of tariff lines featuring specific tariffs from 1790 to 2017, reveals that the share of tariff lines with specific rates was 90 percent in 1790. With the exception of the era spanning the Walker Tariff and Tariff of 1857—which employed an exclusively ad valorem schedule—specific tariffs remained

a prominent form of protection well into the twentieth century, accounting for 45 percent of tariff lines and 60 percent of tariff revenue as late as 1960. Their use then steadily declined to the current share of approximately 10 percent of tariff lines by 2000.

For purposes of comparison, the typical convention is to convert specific and compound tariffs into their ad valorem equivalent, dividing the per-unit dollar amount of the tariff by the price of the imported good. This procedure implies that, in the presence of specific tariffs, the ad valorem equivalent tariff level will depend not only on legislated tariffs but also, mechanically, on price levels. For a given dollar value of a specific tariff, higher prices erode the tariff protection and lower prices enhance it. Across goods and over time, the extent to which tariff protection is influenced by such price fluctuations can be captured by the "specific tariff share," which is the proportion of tariff revenue accounted for by specific tariffs (Greenland and Lopresti 2024). While the effect of changing prices on tariff protection has been documented at the micro-level by Crucini (1994) for a set of goods in the early 1900s and has been recognized as a driver of aggregate ad valorem equivalent tariff rates by Van Cott and Wipf (1983), Irwin (1998), and Greenland, Lake, and Lopresti (2025), the extent to which this has been a primary driver of changes in tariff levels (apparent in Figure 1)—and, by extension, of US openness to trade and globalization—is less well appreciated.

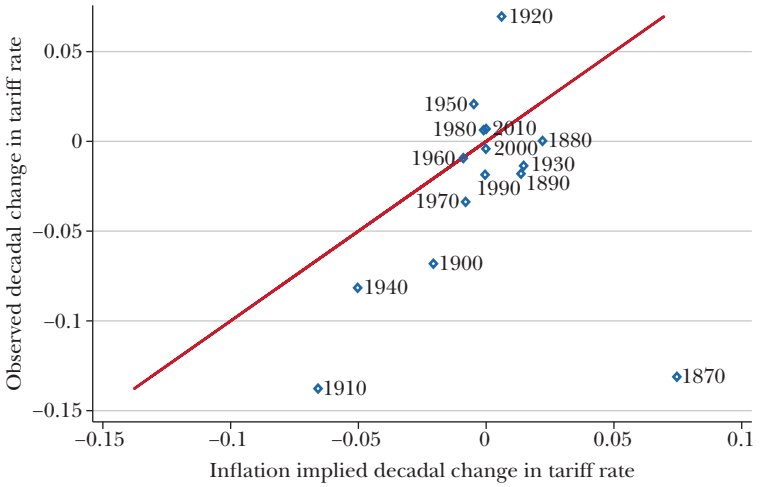
Following Greenland and Lopresti (2024), we characterize the importance of this channel by comparing decadal changes in observed US tariff rates to the changes predicted by subsequent aggregate price movements, given the current share of specific tariffs. This measure captures the magnitude of tariff changes that would have manifested in the absence of statutory rate changes. When prices rise, specific duties are eroded, yielding lower tariff rates; when prices fall, they are enhanced, yielding higher rates.[4] In Figure 7 we present a scatterplot of this relationship along with a 45-degree line.[5] The pattern is striking. Only one decade falls far from the 45-degree line—1870–1880, during which decade-long deflation increased the protective capacity of specific tariffs while legislated tariffs fell in 1870 and 1872, including the movement of coffee and tea onto the duty-free list (Klein and Meissner 2024). Omitting this decade from our sample, a simple regression reveals that 66 percent of the variation in aggregate US tariffs over the last 140 years can be explained by aggregate inflation in the presence of specific tariffs alone. While the broad perception of tariff changes may be that they are primarily statutory, such shifts in fact appear to constitute a minority of variation over time.

[4] Specifically, we regress decadal changes in aggregate AVE tariffs on $-\frac{\Delta CPI_t}{CPI_{t_1}} \times STS_{t_0} \times AVE_{t_0}$, where STS_{t_0} and AVE_{t_0} are the start-of-period aggregate specific tariff share and ad valorem equivalent tariff rate, respectively.

[5] Where necessary data have been converted to decadal equivalents based on neighboring years to facilitate comparability. For example, the 1940 data are not yet incorporated, and so we use the 1939–1950 change scaled by 10/11.

Figure 7

Observed Versus Inflation-Implied Decadal Changes in the US Tariff Rate, 1870–2010



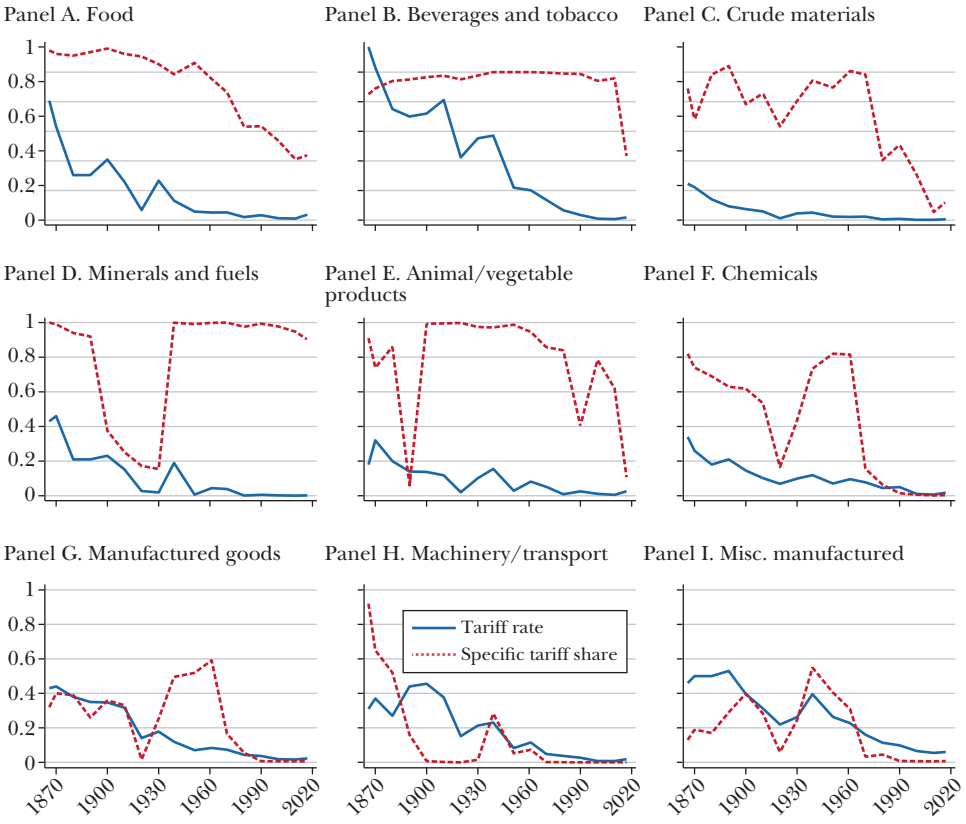
Source: Acosta et al. (2026a).

Note: This figure presents a scatterplot of observed decadal-equivalent tariff changes against those which would have manifest due solely to the inflationary erosion/enhancement of specific tariffs driven by changes in aggregate price levels. Aggregate price changes are calculated using the historical US Consumer Price Index. Specific tariff share and ad valorem equivalent tariff rates are taken from authors' calculations. We also plot a 45 degree line.

A complete accounting of the relative importance of statutory tariff changes over this period would require addressing not only the inflationary channel highlighted above, but also changes in expenditure shares and in the set of goods imported. While such an exercise is beyond the scope of this essay, Greenland, Lake, and Lopresti (2025) provide an exact decomposition of aggregate tariff rates into these channels. In their sample, which spans the 1970s and 1980s, only 30 percent of the variation in aggregate tariff rates can be attributed to statutory rate changes; the remainder reflects the mechanical channel, changes in expenditure shares, and changes in the set of imported goods. They further show that treating all changes in tariffs as statutory can produce welfare estimates that are misleading not only in magnitude but also in direction. The mechanical relationship between price levels and tariff levels thus has important implications both for how we think about the determinants of liberalization and for how we quantify them. The statutory share is likely smaller still the further back one goes, when specific tariffs were more prominent. Serious attempts to evaluate the effects of tariffs on the US over its history must therefore grapple with the centrality of non-policy variation in tariff rates.[6]

[6] In ongoing work, Acosta et al. (2026b) explore the importance of accounting for these different sources of variation in tariff rates when quantifying the effects of tariffs on the macroeconomy.

Figure 8
Ad Valorem Equivalent Tariffs and Specific Tariff Shares by SITC Section



Source: Acosta et al. (2026a)
Note: This figure shows the ad valorem equivalent tariff rate and the specific tariff share of duties for a selection of SITC sections in each decade between 1866 and 2017.

Our tariff-line-level data allow us to compute the specific tariff share of revenue over time for each product between 1866 and 2017, in order to explore sectoral heterogeneity in tariff rates and types. In Figure 8, we plot for a selection of sectors both the sector-level specific tariff share and the ad valorem equivalent tariff. Three patterns emerge. First, consistent with the aggregate patterns shown above, ad valorem equivalent tariffs fell across all sectors over time, albeit at different rates. For some sectors, such as crude materials, protection began low and fell linearly. This stands in contrast to sectors such as Beverages and Tobacco, which had high initial levels of protection and whose cuts were subject to reversion. Second, although ad valorem equivalent tariff levels declined over time across the board, there was persistence in the reliance on specific versus ad valorem tariffs across sectors. That is, while sectors liberalized over time, they did not dramatically alter the mix of

tariff types: raw materials, agricultural goods, food products, and tobacco products remained more reliant on specific tariffs than machinery and manufactured goods. Finally, the prevalence of specific tariffs declined notably toward the end of our sample. Thus, even as ad valorem equivalent tariffs declined, specific tariffs became less important as a share of that falling protection.

What explains sectoral persistence in reliance on specific tariffs, and why has the use of specific tariffs declined over time? At least three explanations have been suggested. First, as proposed in Madsen, Rotemberg, Traiberman, and Wang (2025), the distinction may be an issue of state capacity to enforce the tariff code. Ad valorem tariffs can be subject to under-invoicing if trading parties lie about the value of imported goods. By contrast, specific tariffs distort the relative price of goods, creating additional inefficiencies if not sufficiently targeted. In their framework, while the state prefers a sufficiently targeted specific tariff schedule, which minimizes under-invoicing and distortions, its ability to implement such a schedule is contingent on its capacity. For example, in the presence of an ad valorem tariff, a diamond importer would prefer to minimize duties and convey a diamond as a cheap abrasive rather than an input in, say, expensive jewelry. If the government seeks to legislate distinct specific tariffs on different types of diamonds, it requires more direct state involvement in verification to ensure appropriate application of the tariff code. When the administrative state is small, the government's ability to implement a minimal-distortion specific tariff code is limited. As state capacity grows, it can provide a more differentiated schedule, with targeted rates for each type of good and accompanying enforcement.

A second mechanism to explain both the heterogeneous use of specific tariffs in the cross section and their decline is proposed by Greenland, Lake, Lopresti, and Zhang (2026), who emphasize the volatility-dampening effects of specific tariffs. Because specific tariffs are levied as a fixed dollar amount per unit, they anchor a portion of the tariff-inclusive price independent of world price fluctuations. As a result, specific tariffs generate a price floor for imports and deliver a lower variance of tariff-inclusive prices than ad valorem tariffs of similar magnitude. They argue that the choice of specific versus ad valorem tariffs may reflect a heterogeneous preference for price stability—for example, risk aversion—or heterogeneous need for this form of insurance over time.

A third explanation for cross-sectional heterogeneity relates to legislative inertia and persistence in the structure of the tariff code. Between the 1845 Walker Tariff and the Tariff of 1857, the US tariff code was entirely ad valorem. The Morrill Tariff of 1861 reintroduced specific tariffs. As shown by Greenland and Lopresti (2024), sectoral reliance on specific tariffs under the Morrill Tariff was highly predictive of sectoral reliance on specific tariffs through at least the Tariff Act of 1930. Indeed, the ongoing work of Acosta and Cox (2026) shows that, due to the shifting nature of trade negotiations in the twentieth century, many features of the tariff schedule from the 1930s have persisted until modern times, including sectoral reliance on specific tariffs.

In explaining the overall decline in the aggregate specific tariff share over time, at least two forces are at play. First, rising prices produce a mechanical erosion in the

value of specific tariffs, leading to a decline in both the specific tariff share and the ad valorem equivalent for such tariffs. An inspection of price inflation over the past 60 years suggests this force plays a primary role. This view is also consistent with the interpretation of Irwin (1998), who documents substantial changes in ad valorem equivalent rates between policy changes that are correlated with aggregate price indices. Second, legislation in the late twentieth century converted many specific tariffs to ad valorem. US policymakers reduced their reliance on specific tariffs preceding GATT-mandated statutory tariff cuts—in particular, during the Tokyo and Uruguay Rounds of multilateral trade—as well as during the conversion from the earlier Tariff Schedules of the United States to the Harmonized System in the late 1980s (Greenland, Lake, and Lopresti 2025). During such conversions, policymakers are forced to grapple with the fact that specific tariffs have different effects across trading partners with different export unit values. Thus, when targeting a particular tariff rate cut or establishing a new tariff schedule, policymakers must decide which country's rate to benchmark the new policy against. At least in these cases, a common pattern was to convert specific tariffs to an ad valorem basis prior to phasing them out.

Constructing a Unified Tariff Database

One of the primary aims in constructing the dataset that we describe and rely upon in this article is to provide a consistent series of US tariffs for researchers to employ. Assembling this series required digitizing a range of historical documents containing tariff and import data and developing a consistent classification system, allowing us to track the tariffs on individual products over time. By far the greatest challenge in documenting the historical evolution of tariff protection is the need to develop a consistent product classification. In the earliest years of our data, from 1790 to 1850, creating a panel of product-level tariffs is not possible, as tariff bills were written in plain text, with lists of products that varied considerably from year to year. Similarly, import statistics in these years varied greatly in detail, with some years reporting products narrowly while others record broad aggregates by tariff rate. Madsen, Rotemberg, Traiberman, and Wang (2025) describe these challenges in more detail. With the introduction of the Walker Tariff of 1846, the United States began more systematic tracking of tariff revenue and products. Tariff bills in this era took a more consistent schedule-like structure, though goods at this point were grouped by tariff rate rather than product type. By the 1850s, tariff data become available by country and product and include information on quantity imported in addition to value and tariff rate. Data between this era and 1930 follow a nearly identical structure in editions of the *Foreign Commerce and Navigation of the United States*: goods are listed alphabetically within industries or broad product groups, with detailed descriptions, statutory tariffs, import quantity, value, and implied ad valorem equivalent tariff rate. In cases in which duties are levied on multiple dimensions of a good, two quantities and units may be reported for an individual

Table 1
Data Sources

Year	Tariff data	Import data
1790–1819	Customs legislation	Annual Treasury Reports
1820–1850	Customs legislation	Commerce and Navigation
1850–1946	Foreign Commerce and Navigation of the United States	Foreign Commerce and Navigation of the United States
1947–1962	Schedule A	US Imports of Merchandise for Consumption
1963–1988	Tariff Schedules of the United States	Census Annual Import Data Bank
1980–1987	History of the Tariff Schedules of the United States	Census Annual Import Data Bank
1988–2026	US International Trade Commission	Census International Trade Data

Note: Customs legislation refers to tariff bills that were written into law, for example, the Tariff Act of July 4, 1789. Schedule A refers to "Schedule A Statistical Classification of Imports into the United States." The History of the Tariff Schedules of the United States contains "most favored nation" (MFN) and column-2 rates of duty for the first year in the range, and staged MFN rates for the following years.

product—for example, by the 1920s, tariffs on textiles reflected both weight and the thread-count of the good. Alphabetical lists of tariffs were provided under the Schedule A classification starting in 1903, while the 1920 edition of Schedule A was the first schedule with a hierarchical classification system similar to those used today. Schedule A was used until 1962, after which the United States transitioned to the Tariff Schedules of the United States, which was used until 1988, at which point the modern-day Harmonized Tariff System was adopted. Even within one such era, products are often "split," or "merged" over time, requiring detailed tracking of individual products across regimes.

It is difficult to track products across these different classification systems. Even within classification regimes, there were many changes in the codes used to classify products from one year to the next. For example, Acosta and Cox (2025) document that within the three classification regimes that existed in the 1930–1989 period, there were over 43,000 product code changes, over two-thirds of which had to be mapped manually by comparing tariff schedules from one year to the next. To provide a consistent classification system across the full dataset, we have manually mapped products in each year to the SITC (Standard International Trade Classification), Revision 2 classification. In more recent years, processes for this mapping already exist—for example, from the Harmonized Tariff System to the SITC—but in earlier years we rely on the text descriptions in the tariff data to find the appropriate SITC code.

We chose the SITC (Revision 2) classification system due to several attractive features. First, it overlaps with the modern Harmonized System classification and is based on post-1925 classification systems, which were the first to provide numbered product groups. Second, a mapping between the SITC and the earlier Tariff Schedules of the US classifications have previously been developed in Feenstra (1996).

Finally, the SITC defines products at a relatively disaggregated five-digit level, providing a better match than more recent systems to data preceding 1925.

We are typically able to map the historical tariff data between the three- and five-digit level in the SITC system. As one example, under the Tariff Act of 1930, 5,000 unique tariff lines can be mapped to 450 four-digit SITC industries. Such disaggregation allows for distinction at a product level. For example, the one-digit SITC code "0" corresponds to Food and Live Animals Chiefly for Food. The second digit differentiates among live animals (00) and meat and meat preparations (01). The third digit delineates means of processing, distinguishing fresh, chilled, or frozen meats (011) from salted, brined, dried, or smoked (012). The fourth digit distinguishes between prepared meats by type of animal, with 0121 covering swine (bacon, ham, and so on) for example. Naturally, the level of disaggregation varies across SITC categories, with an interquartile range of 13–69 tariff lines within four-digit SITC codes during this era. Textiles and their manufactures are by far the most differentiated products, with even narrowly defined SITC categories exhibiting different rates based on material, purpose, composition, whether hemmed or not, coloring and dyeing, and thread count. Such variation, while present in our data, maps to 48 four-digit level products within the SITC 65 code.

Importantly, data for the years after 1848 typically cover tariffs among traded goods only: products for which tariffs were sufficiently high to eliminate imports, or minor products with small import values, may not have been reported in the statistical summaries on which we rely. While the dataset thus may not fully reflect the restrictiveness of the tariff schedule, our focus on traded goods was intentional.[7] When tariffs are levied on a per-unit basis, calculation of a tariff rate requires conversion to an ad valorem equivalent rate. This requires the price of the good under consideration, which is only observed for traded goods. Given the prevalence of these tariffs throughout US history, we chose to focus exclusively on traded products throughout our sample.

New Data, New Directions

In this concluding section, we attempt to answer some primary questions on the role of tariffs throughout US history based on our ongoing research. In particular, we focus on recent studies of the effects of tariffs on productivity, structural changes, novel identification strategies, measurement issues, and important caveats for interpreting the distributional consequences of tariff policy over time.

Why Specific Tariffs Matter for Empiricists

As we emphasize above, the heterogeneous structure of the tariff code—in particular, the choice of ad valorem versus specific tariffs—is a key feature of our

[7] For the period 1866 to 1900, the share of total recorded imports in our data compared to the official statistics averages 98.4 percent, suggesting that the nontraded share is negligible.

data. We begin by emphasizing why the ability to distinguish between specific and ad valorem rates can be crucial for empiricists attempting to measure the causal effects of trade policy. In general, tariff levels are endogenous to economic outcomes both statutorily and mechanically. Statutorily, politically manipulated tariffs may reflect the demand for protection (Grossman and Helpman 1994). In empirical work, this endogeneity concern is typically addressed by exploiting plausibly exogenous changes in trade policy or with an instrumental variable strategy. The presence of specific tariffs implies that tariffs are also related to economic conditions due to the mechanical effect of prices on the ad valorem equivalent.

To the extent that the price movements that induce changes in tariff protection are unanticipated, however, they are not subject to the same political economy concerns associated with statutory tariff changes. In this way, as Greenland and Lopresti (2024) note, the stochastic protection offered by specific tariffs can be beneficial to researchers. Of course, the ad valorem equivalent of specific tariffs may still be endogenous to domestic outcomes if the underlying price changes are driven by domestic economic conditions.[8] To circumvent this concern, Greenland and Lopresti (2024) exploit price variation at a higher level of aggregation, such that price-induced tariff changes are not driven by observation-level demand or supply shocks. They demonstrate the usefulness of this identification strategy between 1900 and 1940 and find that variation driven by price movements in the presence of specific tariffs accounts for nearly 40 percent of the industry-level changes in ad valorem equivalent tariffs during these years. This variation is highly predictive of import growth and can be used to explore the effects of tariff protection on domestic labor-market outcomes in linked US Census data. The authors find that tariffs played an important role in shaping the US transition from an agricultural- to a manufacturing-based economy.

Following this identification strategy, Klein and Meissner (2024) explore the relationship between tariffs, labor productivity, employment, and gross output in the manufacturing sector between 1870 and 1910. Using disaggregated state-level and sector-level observations from the Census of Manufactures, they find that price-driven changes in tariff protection led to higher relative employment and output, but smaller average firm size and lower labor productivity. However, in some of the most concentrated industries of the time—such as tobacco products and meat packing—tariffs were associated with declines in output and employment. They interpret these findings as consistent with the idea that higher tariffs generally restricted domestic and foreign competition, lowered innovation, and raised output prices for US consumers. Given the prevalence of specific tariffs documented above,

[8]This concern is not unique to this identification strategy. For instance, the well-known paper by Autor, Dorn, and Hanson (2013), on the subject of how rising imports from China affected local US labor markets, also requires variation not driven by domestic outcomes. However, it should be broadly noted that evidence on the effects of changes in trade policy is often framed through the lens of "import competition" in reduced form studies, and while such studies are informative, they are subject to the well-known missing intercept problem. Evidence of tariff cuts yielding worse relative economic outcomes can be misinterpreted as a statement about level effects by the public.

this variation is likely to prove useful to researchers in various settings. As specific tariffs are not unique to US trade policy, this strategy is also promising in other countries and eras.

At the same time, specific tariffs may threaten identification if researchers use ad valorem equivalent tariff rates without understanding the extent to which measured tariff changes are driven by price variation rather than statutory rate changes. Greenland, Lake, and Lopresti (2025) explore this issue in the years surrounding the Tokyo Round of the GATT and show that naive handling of tariff data can lead to biased and even incorrectly signed estimates. Reductions in ad valorem equivalent tariffs correspond to increasing imports if they are driven by reduced statutory rates. However, if falling ad valorem equivalent tariffs are driven by the inflationary erosion of specific tariffs, lower measured tariffs may be associated with falling imports. That is, because there is a mechanical inverse relationship between price levels and tariff levels, changes in ad valorem equivalent rates may be endogenous to outcomes of interest precisely because they reflect the same price movements. Interpreting the relationship between changes in ad valorem equivalent tariffs and economic outcomes of interest requires understanding whether the changes are driven by statutory or inflationary channels. Our comprehensive database has all of the data necessary to isolate these confounding effects.

The Economic Effects and Political Economy of Tariffs

Changes in US tariffs over time have not typically been uniform across sectors, but have often reflected the special interests of the policymakers. This remained true even as GATT-mandated liberalizations became the dominant form of liberalization throughout the second half of the twentieth century. These liberalizations often targeted average reductions in tariffs, but allowed countries opportunities to offer greater than mandated cuts in some sectors while sparing others from significant liberalization. Granular historical data on tariffs can be used to assess the performance of favored industries and the political attitudes toward trade in more exposed locations. While Acosta and Cox (2025) have emphasized how these features altered the regressivity of the US tariff code, here we suggest a few cases in which data on the level and structure of tariffs over time could enable innovative research on classic questions regarding the political economy of tariff setting.

First, the current wave of antiglobalization sentiment, coinciding with the rise of China as an economic superpower, has raised questions about the roles played by surging imports and US distributional considerations in predicting support for protectionism. Insights might be gained from an historical accounting of the direct effects of previous import surges on the US economy. There has been some recent work on import responses to country-specific liberalizations using disaggregated data, but the effects on other domestic outcomes of prior rounds of liberalization have gone largely unstudied. Recent work addressing this gap in the literature includes Alessandria et al. (2025), who revisit a major liberalization vis-à-vis China in 1980, while Greenland, Lake, and Lopresti (2026) explore the effects of the Tokyo Round of the GATT on inequality.

Attitudes toward tariffs could also be shaped by the extent to which macroeconomic conditions such as the degree of exchange rate flexibility facilitate adjustment to trade shocks. Both the 1971 "Nixon shock" (Irwin 2013) and 1985 Plaza Accord to adjust exchange rates put a sharp focus on the interaction between exchange rate policy and tariff policy. In both cases, the threat of rising trade barriers was used to offset what policymakers viewed as a problem induced by an overvalued US dollar, which differentially harmed US manufacturing output and exports. Tariff changes across history, including the eras in which the US dollar did not serve as a global reserve currency, may offer unique opportunities to evaluate interactions between tariffs and exchange rates. These data offer opportunities to explore how exchange rates and other macroeconomic policy tools work in complement with or substitute for tariff policy.

Earlier tariff episodes may also cast light on tariffs as a form of industrial policy—that is, whether temporary and focused increases in tariffs served to bolster particular US industries. In theory, such tariffs could be unwound once the industries are strong enough to face global competition. There is very little evidence, however, on how quickly such reindustrialization might or does take place. While tariff cuts do seem to be associated with relative declines in manufacturing (Pierce and Schott 2016), it need not be the case that tariff hikes induce firm entry—especially in the face of uncertainty over the extent of commitment to high tariff levels in the future. A related concern is that the use of tariffs for industrial policy or other purposes may create demand for future protection. While the canonical view of tariffs often frames policy as an equilibrium outcome (Grossman and Helpman 1994), it seems reasonable to believe that if tariffs lead to a reorganization of supply chains to support domestic manufacturing, for example, then suppliers will have a vested interest in maintaining tariffs as part of a new status quo.

While we have focused largely on our own work on these topics, we hope that other researchers will take advantage of the data described in this article when they are made available in order to better inform scholars, policymakers, and the lay public of the role that tariffs have played in the development of the US economy and their effects on US workers, consumers, prices, and firms. We hope, too, that such information may inform our views of what optimal policy looks like moving forward. Whether the post-2025 moment proves to be a genuinely new regime or a reversion to older ideas, the answer will turn on the same political and economic fundamentals that have shaped US tariff policy since 1789—fundamentals that the data described here are well-positioned to bring into view.

■ *Acosta, Cox, Greenland, Lopresti, and Meissner benefited from support under NSF Award #2314696. Greenland and Lopresti also thank the Upjohn institute for support via the ECRA program. Meissner gratefully acknowledges support from the UC Davis Levine Funds. We would also like to acknowledge the contributions of coauthors: Alexander Klein, James Lake, Erik Madsen, and Shizhuo Wang. We are grateful for guidance from the JEP editorial board. We would like to thank seminar participants at the NSF conference on US Tariff Policy hosted at NC State University.*

References

Acosta, Miguel, and Lydia Cox. 2025. "The Regressive Nature of the US Tariff Code: Origins and Implications." Unpublished.

Acosta, Miguel, and Lydia Cox. 2026. "Sticky Tariffs." Unpublished.

Acosta, Miguel, Lydia Cox, Andrew Greenland, John Lopresti, Christopher M. Meissner, Martin Rotemberg, and Sharon Traiberman. 2026a. *Data and Code for: "US Tariff Policy Since 1789."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E250176V1.

Acosta, Miguel, Lydia Cox, Andrew Greenland, James Lake, John Lopresti, and Stefano Lord-Medrano. 2026. "The Aggregate(d) Effects of Tariff Policy." Unpublished.

Alessandria, George, Shafaat Yar Khan, Armen Khederlarian, Kim J. Ruhl, and Joseph B. Steinberg. 2025. "Trade Policy Dynamics: Evidence from 60 Years of US China Trade." *Journal of Political Economy* 133 (3): 713–49.

Amiti, Mary, Stephen J. Redding, and David E. Weinstein. 2019. "The Impact of the 2018 Tariffs on Prices and Welfare." *Journal of Economic Perspectives* 33 (4): 187–210.

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

Bagwell, Kyle, and Robert W. Staiger. 1999. "An Economic Theory of GATT." *American Economic Review* 89 (1): 215–48.

Bagwell, Kyle, and Robert W. Staiger. 2002. *The Economics of the World Trading System*. MIT Press.

Bagwell, Kyle, Robert W. Staiger, and Ali Yurukoglu. 2020. "Multilateral Trade Bargaining: A First Look at the GATT Bargaining Records." *American Economic Journal: Applied Economics* 12 (3): 72–105.

Caliendo, Lorenzo, and Fernando Parro. 2015. "Estimates of the Trade and Welfare Effects of NAFTA." *Review of Economic Studies* 82 (1): 1–44.

Congressional Globe. 1838. "Defalcation of Samuel Swartwout." https://www.congress.gov/congressional-globe/appendix-page-headings/25th-congress/Defalcation%20of%20Samuel%20Swatwout/2111.

Crucini, Mario J. 1994. "Sources of Variation in Real Tariff Rates: The United States, 1900–1940." *American Economic Review* 84 (3): 732–43.

Fajgelbaum, Pablo D., Penelope K. Goldberg, Patrick J. Kennedy, and Amit K. Khandelwal. 2020. "The Return to Protectionism." Quarterly Journal of Economics 135 (1): 1–55.

Feenstra, Robert C. 1996. "US Imports, 1972–1994: Data and Concordances." NBER Working Paper 5515.

Feng, Emily. 2025. "Trump Has Imposed a Lot of Tariffs. But Here's Why Collecting Them Can Be Hard." *NPR*, May 20. https://www.npr.org/2025/05/20/nx-s1-5387330/trump-tariffs-china-collect-manufacturing-revenue.

Goss, John Dean. 1897. *The History of Tariff Administration in the United States: From Colonial Times to the McKinley Administrative Bill*. Columbia University Press.

Grant, Matthew. 2023. "Where Do the Data Come From? Endogenous Classification in Administrative Data." Unpublished.

Greenland, Andrew, James Lake, and John Lopresti. 2025. "AVE Tariffs are a Misleading Measure of Tariff Policy." NBER Working Paper 34429.

Greenland, Andrew, James Lake, and John Lopresti. 2026. "Trade and U.S. Inequality in the Tokyo Round." *Review of Economics and Statistics*. https://doi.org/10.1162/REST.a.1774.

Greenland, Andrew, James Lake, John Lopresti, and Z. Zhang. 2026. "Tariff Policy: First and Second Moments." Unpublished.

Greenland, Andrew, and John Lopresti. 2024. "Trade Policy as an Exogenous Shock: Focusing on the Specifics." NBER Working Paper 33127.

Grossman, Gene M., and Elhanan Helpman. 1994. "Protection for Sale." *American Economic Review* 84 (4): 833–50.

Hakobyan, Shushanik, and John McLaren. 2016. "Looking for Local Labor Market Effects of NAFTA." *Review of Economics and Statistics* 98 (4): 728–41.

Huger, B. 1816. "Unsettled Balances." American State Papers: Finance, 14th Congress, 1st Session, No. 486. https://archive.org/details/americanstatepap03unit/page/123/mode/1up.

Ikenson, Daniel J. 2018. "*Clashing over Commerce: A History of US Trade Policy* by Douglas A. Irwin." https://www.cato.org/cato-journal/spring/summer-2018/clashing-over-commerce-history-us-trade-policy-douglas-irwin.

Irwin, Douglas A. 1998. "Changes in U.S. Tariffs: The Role of Import Prices and Commercial Policies." *American Economic Review* 88 (4): 1015–26.

Irwin, Douglas A. 2013. "The Nixon Shock after Forty Years: The Import Surcharge Revisited." *World Trade Review* 12 (1): 29–56.

Irwin, Douglas A. 2017. *Clashing over Commerce: A History of US Trade Policy*. University of Chicago Press.

Johnston, Louis, and Samuel H. Williamson. 2026. "The Annual Real and Nominal GDP for the United States, 1790–1928." *MeasuringWorth*.

Jones, Andrew A. 1835. *Jones's Digest: Being a Particular and Detailed Account of the Duties Performed by the Various Officers Belonging to the Custom-House Departments of the United States, but More Especially of Those Attached to the District of New York*. G. F. Hopkins and Son.

Jones, Andrew A. 1854. *The Revenue Book: Containing the New Tariff of 1846, Together with the Tariff of 1842*. Geo. H. Bell.

Klein, Alexander, and Christopher M. Meissner. 2024. "Did Tariffs Make American Manufacturing Great? New Evidence from the Gilded Age." NBER Working Paper 33100.

Krueger, Anne O. 1974. "The Political Economy of the Rent Seeking Society." *American Economic Review* 64 (3): 291–303.

Madsen, Erik, Martin Rotemberg, Sharon Traiberman, and Shizhuo Wang. 2025. "Tariffs and State Capacity: A Specific Example." NBER Working Paper 34143.

Maggi, Giovanni. 1999. "The Role of Multilateral Institutions in International Trade Cooperation." *American Economic Review* 89 (1): 190–214.

Ossa, Ralph. 2011. "A 'New Trade' Theory of GATT/WTO Negotiations." *Journal of Political Economy* 119 (1): 122–52.

Pierce, Justin R., and Peter K. Schott. 2016. "The Surprisingly Swift Decline of US Manufacturing Employment." *American Economic Review* 106 (7): 1632–62.

Rao, Gautham. 2016. *National Duties: Custom Houses and the Making of the American State*. University of Chicago Press.

Romalis, John. 2007. "NAFTA's and CUSFTA's Impact on International Trade." *Review of Economics and Statistics* 89 (3): 416–35.

Roosevelt, Franklin D. 1934. "Message to Congress Requesting Authority Regarding Foreign Trade." March 2. https://www.presidency.ucsb.edu/documents/message-congress-requesting-authority-regarding-foreign-trade.

Taussig, Frank W. 1910. *The Tariff History of the United States*. G.P. Putnam's Sons, Knickerbocker Press.

Trump, Donald J. 2025. "Remarks Announcing Additional United States Tariff Actions on Foreign Imports." April 2. https://www.presidency.ucsb.edu/documents/remarks-announcing-additional-united-states-tariff-actions-foreign-imports.

US Congress. 1820. *An Act to Provide for obtaining Accurate Statements of the Foreign Commerce of the United States*. 16th Cong., 1st Sess., Ch. 11, 3 Stat. 541.

US Department of Commerce. 1930. *Statistical Classification of Imports into the United States*. US Government Printing Office.

US Department of Commerce. 1937. *Statistical Classification of Imports into the United States*. US Government Printing Office.

US Department of State. 1816–1850. *Official Register of the United States*. US Government Printing Office.

US Department of the Treasury. 1980. *Annual Report of the Secretary of the Treasury on the State of the Finances, Fiscal Year 1980: Statistical Appendix*. Fiscal Year US Government Printing Office.

US Tariff Commission. 1939. *Twenty-Third Annual Report of The United States Tariff Commission*. US Government Printing Office.

US Treasury Department. 1839. *Defalcation of Samuel Swartwout*. House of Representatives, 25th Congress, 3rd Session, Document No. 69.

Van Cott, T. Norman, and Larry J. Wipf. 1983. "Tariff Reduction via Inflation: U.S. Specific Tariffs 1972–1979." *Review of World Economics* 119 (4): 724–33.

Waugh, Michael E. 2026. "Trade War Tracker." https://www.tradewartracker.com/.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 73–100

Labor Market Responses to Tariffs: Frictions, Dynamics, and Policy Responses

Rafael Dix-Carneiro and Brian K. Kovak

In recent decades, many countries around the world have seen a backlash against international trade, reflected in rising anti-globalization political sentiment, shifts in voting behavior, and substantial increases in protectionist policies (Goldberg and Reed 2023). Colantone, Ottaviano, and Stanig (2022) survey the evidence on the drivers of this backlash and identify trade-induced inequality across social groups and regions as the central force driving these changes. These gradual shifts became acute in 2025, when the Trump administration began implementing a broad increase in US import tariffs. The administration justified its tariff policies with a wide range of rationales, including reducing trade deficits, countering allegedly unfair trade practices, addressing national security concerns, and generating tariff revenue. But perhaps its most prominent rationale was the claim that import competition—particularly from China—harms US workers by reducing their earnings and employment.

In response to this concern, this article provides an introduction to the evidence and modeling frameworks contemporary economists use to understand the effects of trade policy and import competition on labor markets. We start

■ *Rafael Dix-Carneiro is Professor of Economics, Duke University, Durham, North Carolina. He is also a Research Associate, Bureau for Research and Economic Analysis of Development (BREAD), Princeton, New Jersey. Brian K. Kovak is Professor of Economics and Public Policy, Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, Pennsylvania. He is also a Research Associate, Institute of Labor Economics (IZA), Luxembourg Institute of Socio-Economic Research, Alzette, Luxembourg. Both authors are Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are rafael.dix.carneiro@duke.edu and bkovak@cmu.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20251489.

by discussing some of the classic workhorse models of trade, which rely on stark assumptions about the mobility of workers and capital. One class of these models assumes factors are perfectly mobile across sectors and regions, an approach typically used to characterize long-run outcomes. In this spirit, we briefly review a result that many readers will find familiar: the Stolper-Samuelson theorem in the context of the Heckscher-Ohlin model, which focuses on the long run by assuming perfectly mobile factors of production. At the other extreme are frameworks in which one or more factors of production are completely immobile, an assumption commonly employed to study short-run impacts. In particular, we discuss the specific-factors model, in which some factors can only be used in a single industry. As we will discuss, this model provides a central building block for the empirical literature studying the effects of national trade policy changes on local labor markets.

Neither extreme, however, captures the gradual and uneven adjustment processes that shape labor market outcomes over time. Central concerns of policymakers and the public lie precisely in this intermediate horizon. Even if trade liberalization generates substantial gains in the long run, how quickly are these gains realized? Which workers, regions, sectors, and occupations bear the losses during the transition? How do institutional features such as labor market regulations and their enforcement shape these outcomes? And what role do trade imbalances play in the adjustment process to trade shocks?

Over the past 15 years, a large and influential literature in international trade and labor economics has studied these questions from both empirical and theoretical perspectives. We begin by discussing recent empirical studies that credibly identify the causal labor market effects of trade shocks. A central finding is that trade liberalization and import competition can generate sizable and highly localized labor market disruptions. Specifically, employment and earnings losses tend to be concentrated in particular regions, sectors, and occupations, and adjustment is often slow, with effects extending beyond directly affected workers to entire local communities. We then turn to recent evidence on the labor market consequences of increased tariff protection and examine whether it delivers the employment gains often claimed by its proponents.

The findings from this empirical literature have, in turn, motivated the development of quantitative trade and labor market models that incorporate these adjustment patterns and are better suited to informing policymakers about the speed of adjustment, which workers lose during the transition, and the policies that can mitigate losses. In particular, the literature has increasingly moved from static to dynamic frameworks, thus emphasizing mobility frictions across regions, sectors, and occupations, along with worker heterogeneity and labor market regulations.

The insights emerging from this recent literature have important policy implications. In an environment of rising skepticism toward globalization, governments pursuing a pro-trade agenda face the challenge of designing complementary policies that mitigate adjustment costs and support adversely affected workers and

regions. Thus, we conclude by discussing policies aimed at smoothing worker transitions and compensating those who lose as a result of trade policy changes.

Long-Run Effects in a Frictionless World

If one asks advanced undergraduate economics students to predict the effect of changes in trade policy on factor prices, they are likely to cite the Stolper-Samuelson (1941) theorem in the context of the canonical two-industry and two-factor Heckscher-Ohlin model (throughout, we use "factor" and "input" interchangeably). The theorem predicts an increase in the real returns to the factor used intensively in producing the good whose relative price increased, while the other factor's real returns decrease, irrespective of the industry employing those factors. To be more specific, consider an economy with two products, A and B, with prices P_A and P_B. Both goods are produced by perfectly competitive, profit-maximizing firms using factors X and Y, whose prices are w_X and w_Y. We assume that both factors are perfectly mobile across industries and that product A is intensive in the use of factor X, implying that product B is relatively intensive in the use of factor Y.[1] Let hats represent proportional changes, such that $\hat{x} \equiv \Delta x/x$. If the change in trade policy leads to product price changes such that $\hat{P}_B > \hat{P}_A$, the Stolper-Samuelson result is that $\hat{w}_Y > \hat{P}_B > \hat{P}_A > \hat{w}_X$. In other words, the price of the intensive factor in the favorably affected industry increases relative to both prices, while the price of the other factor falls relative to both prices, a result Jones (1965) described as "magnification." Samuelson (1987) recalls experiencing "excited surprise" upon discovering this result in part because it differed from partial equilibrium models, in which output price changes tend to drive changes of smaller magnitude in their associated factor prices.

Despite the appeal of the Stolper-Samuelson theorem, its applicability has limits. It yields sharp predictions in a two-by-two model, but when generalizing to more realistic settings with multiple products and factors, it often leads to indeterminacies or requires unintuitive technological restrictions to recover sharp predictions in empirically relevant contexts (as discussed in Deardorff (1994) and its associated volume). For example, the commonly-cited "friends-and-enemies" formulation states that, irrespective of the number of products and factors, a change in the price of one product will raise the real return to some factor and reduce the real return to some other factor (Ethier 1974; Kemp and Wan 1976; Jones and Scheinkman 1977). While this generalization has an intuitive plausibility, the result is of limited practical use because it does not specify which factors will win and lose when a price change occurs nor does it provide information on how other factors will be affected.

The two-by-two model's predictions are also seemingly contradicted by salient empirical examples. For example, many developing countries that substantially

[1] Further assume that both production functions are linearly homogeneous, twice differentiable, strictly quasi-concave, and have strictly positive marginal products.

reduced trade barriers between the late 1970s and the early 1990s experienced subsequent increases in inequality (Goldberg and Pavcnik 2007). With skilled and unskilled labor as inputs, and assuming that developing countries are relatively abundant in unskilled labor, the Stolper-Samuelson theorem predicts the opposite, an increase in the relative wage of locally abundant unskilled workers and an ensuing reduction in inequality. This apparent contradiction motivated researchers to explore alternative mechanisms linking trade and inequality, as discussed in detail in Dix-Carneiro and Kovak (2025).

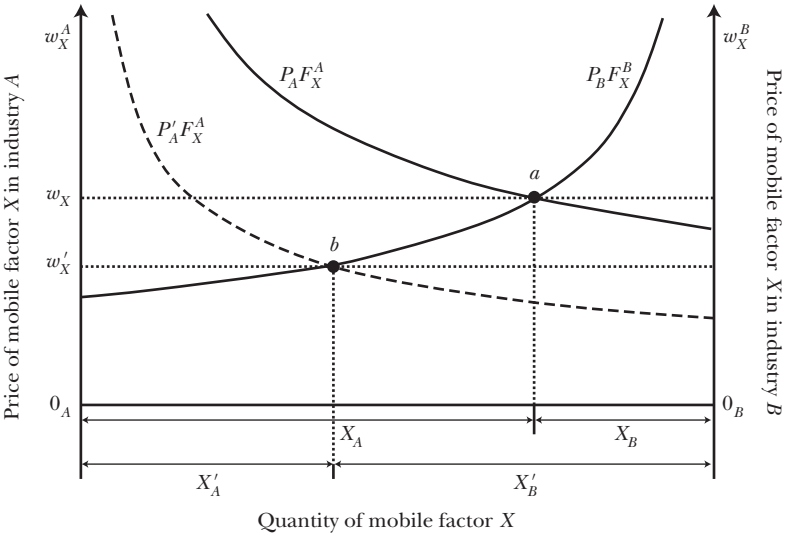
The Stolper-Samuelson theorem also applies only to a world without frictions, in which all factors of production are able to transition costlessly between sectors. While this feature might be relevant in the long run, many factors face short-run costs when switching industries, including the cost of redeploying a machine in a different factory or allowing a worker to accumulate skills specific to a new industry. Moreover, in the frictionless setting, a worker or capital owner's industry is irrelevant to their preferences over trade policies, in which case all workers should oppose a tariff on the capital-intensive good irrespective of their current industry of employment. Yet, in reality, workers and owners often align in their lobbying efforts for protection of their industry, without regard to factor intensities (Magee 1980; Beaulieu 2002; Mayda and Rodrik 2005). Even Stolper and Samuelson (1941) acknowledge: "Nobody, of course, ever denied that the workers employed in the particular industry which loses a tariff could be hurt in the short-run . . . ." These observations led researchers to consider alternative frameworks incorporating factor-mobility frictions.

Short-Run Effects and Industry-Specific Factors

The simplest and most extreme version of mobility frictions assumes that all factors are entirely industry specific—that is, factors cannot switch industries at all. In a very short-run model where all factors are industry specific, the implications of price changes induced by trade policy are trivial; with competitive factor markets and perfectly inelastic factor supply, a change in the price of a given product results in a proportional change in the price of its associated factors. While such a model might be relevant when inputs are highly specialized or in the very short run, the literature has focused on a model allowing for some factor mobility, in which each industry has a single industry-specific factor and all industries share a common mobile factor. This setup is known as the Ricardo-Viner or specific-factors model.

Consider an economy identical to the one discussed in the previous section except that the production of product A uses a mobile factor X and an industry-specific factor Y_A, while production of B uses X and an industry-specific factor Y_B. While Y_A and Y_B are used exclusively in their respective industries, X is freely mobile, so its factor payment is equalized across industries. We can visualize the equilibrium using the diagram in Figure 1, which first appeared in Mussa (1974). The length of the horizontal axis represents the amount of the mobile

Figure 1
Specific-Factors Diagram



Source: Authors' creation, based on Mussa (1974).
Note: This figure visualizes equilibrium in the specific-factors model. The length of the horizontal axis represents the amount of the mobile factor, *X*. The left vertical axis plots the price of *X* in industry *A*, w_X^A, while the right vertical axis plots the same for industry *B*, w_X^B. The solid curves labeled $P_A F_X^A$ and $P_B F_X^B$ are the price times marginal product of *X* in the relevant industry. Point *a* represents the initial equilibrium, with mobile factor price w_X and mobile factor allocation X_A and X_B. Point *b* represents the equilibrium after a reduction in the price of product *A* from P_A to P'_A, with mobile factor price w'_X and mobile factor allocation X'_A and X'_B.

factor, *X*. The left vertical axis plots the price of *X* in industry *A*, w_X^A, while the right vertical axis plots the same for industry *B*, w_X^B. The share of *X* employed in industry *A* increases as one moves to the right from 0_A. The solid curves labeled $P_A F_X^A$ and $P_B F_X^B$ are the price times marginal product of *X* in the relevant industry, which are decreasing in the amount of *X* employed in the industry because the quantity of each specific factor is fixed. Because factor markets are competitive and *X* is freely mobile between industries, the values of marginal product are equal across industries and equal to the price of *X*. Because the length of the *x*-axis reflects the total quantity of *X* in the economy, the intersection of the two value of marginal product curves at point *a* determines the factor price w_X and the allocation of *X* across the two industries, X_A and X_B.

Given this setup, imagine that the country implements a tariff cut that reduces the price of product *A* to $P'_A < P_A$. This change shifts the value of marginal product in industry *A* downward, leading to a new equilibrium at point *b*, with a reduction in the price of *X* to $w'_X < w_X$ and a reallocation of the mobile factor away from industry *A* and toward *B*. How does this change affect the real returns to each factor? For the mobile factor, the answer depends upon the preferences of its

owners. Because w_X fell in nominal terms, its purchasing power falls for product B, whose price was unchanged. However, the decline in w_X is smaller than the reduction in P_A. This implies an increase in purchasing power in terms of product A, so the net effect of the price change depends on X owners' preferences for A versus B.

Payments to the industry-specific factors can be seen on the figure as the area below the associated value of marginal product curve and above the mobile factor price.[2] The decline in P_A clearly increases payments to Y_B and also reduces payments to Y_A, because w_X falls less than P_A. This immediately implies an increase in the real return to Y_B and a decline in Y_A's purchasing power in terms of B. Y_A's purchasing power also falls in terms of A; the reduction in X_A implies a decline in F_Y^A, because X and Y are complements, and therefore w_Y^A falls relative to P_A, because $w_Y^A = P_A F_Y^A$. Hence, the real return to industry-A specific factor falls. To summarize these findings in a situation where $\hat{P}_B > \hat{P}_A$, the specific-factors model implies that $\hat{w}_Y^B > \hat{P}_B > \hat{w}_X > \hat{P}_A > \hat{w}_Y^A$. In other words, the specific factor in the industry experiencing a relative price increase gains, the other specific factor loses, and the effect on the mobile factor is ambiguous, depending on its owners' preferences. Consistent with the friends-and-enemies generalization, the specific factors experience magnification effects similar to the Stolper-Samuelson theorem in general equilibrium, while the mobile factor experiences an effect like in a partial equilibrium model (Jones 1971).

When labor is the mobile factor and capital the specific factor, workers and owners will in some cases agree to support protection for their current industry. Capital owners in the industry will benefit unambiguously, and workers will also benefit if their consumption is weighted toward the other industry. This pattern is in sharp contrast with the Stolper-Samuelson theorem, in which the interests of capital and labor will always be in conflict, because any product price change will help one factor and harm the other. Mayer (1974) and Mussa (1974) interpret the specific-factors model as reflecting the short run and the Heckscher-Ohlin model as the long run, emphasizing the differences in each factor's support for or opposition to protection in the short- versus long-run. Note, however, that despite the presence of winners and losers, both the Heckscher-Ohlin and specific-factors models imply that trade generates aggregate gains at the country level—that is, trade increases the overall size of the economic pie in all trading partners. In principle, because the aggregate gains exceed the losses, the winners from trade could compensate the losers while still remaining better off themselves, so that everyone would be weakly better off relative to a situation with less trade openness. We return to this idea toward the end of the paper, where we discuss policies aimed at smoothing the adjustment costs associated with trade liberalization.

[2] Consider the payment to industry-A specific factor in the initial equilibrium at point a. Total revenue in industry A is the area under the value of marginal product curve, $P_A F_X^A$, as integrating the marginal product yields total output. The total payments to the mobile factor in industry A are $w_X X_A$, corresponding to the rectangular area below and to the left of point a. With perfect competition and zero profits, total payments to Y_A are total revenue minus payments to the mobile factor—that is, the area below $P_A F_X^A$ and above w_X.

Along with the benefit of allowing analysts to consider effects in the short run or in contexts where productive inputs are inherently industry-specific, as with some machinery and natural resources, the specific-factors model also admits a straightforward generalization to many industries, each of which is assumed to have an industry-specific factor and to share a common mobile factor (Jones 1975). This model avoids the indeterminacies of the more general Heckscher-Ohlin model and largely maintains the results just worked out in the two-industry case. In this many-industry specific-factors model, the mobile factor's price change is a weighted average of price changes across industries, where industries employing a larger share of the mobile factor receive greater weight. In the following section, we show how studies of the effects of trade on local labor markets have capitalized on the specific-factors model's ability to generalize to settings with many industries, facilitating empirical analysis using fine-grained industry data.

Static Effects of Trade on Local Labor Markets

In the 1990s, as detailed data on product-level trade flows and worker-level labor market outcomes became more widely available, researchers began exploring new dimensions of how trade policy might potentially affect labor markets. Many papers studied the effects of changes in trade policy or other measures of import competition on wages at the industry level, motivated by the idea that labor is industry-specific, at least in the short run, and therefore tied to the prospects of the employing industry. Prominent early examples include Revenga (1992) on the United States, Revenga (1997) on Mexico, and Attanasio, Goldberg, and Pavcnik (2004) and Goldberg and Pavcnik (2005) on Colombia, all of which find significant relative reductions in industry wage premia when facing larger reductions in trade protection, as predicted if labor is partly industry specific. This cross-industry approach raises an important interpretation issue: it can only identify *relative* effects between industries facing different changes in trade protection, not *aggregate* consequences. For example, the papers just mentioned are consistent with the interpretation that trade liberalization raised wages across the board, but did so less in industries facing larger tariff cuts (for a detailed discussion, see Dix-Carneiro and Kovak 2025, Section 3.1.3).

Because industries tend to be concentrated in particular locations within countries and workers face costs of moving geographically, these industry-level effects of changes in *national* trade policy may lead to different effects across *local* labor markets. In a pair of seminal papers, Topalova (2007, 2010) studied the effects of trade liberalization in India on regional poverty through the lens of a two-industry specific-factors model, in which labor is industry specific and immobile across locations. Her approach related changes in regional poverty rates to an employment-weighted average of tariff changes, placing more weight on industries employing a larger share of regional labor. In this framework, regions face different local tariff shocks due to differences in industry mix across locations and

differences in tariff changes across industries. Both papers find that Indian regions whose industries faced larger tariff cuts experienced slower reductions in poverty than other regions.

Topalova's insight of combining trade policy variation across industries with regional variation in industry mix using a weighted average has proven particularly influential in the subsequent two decades because of its intuitive appeal and empirical relevance to observed regional outcomes. Kovak (2013) develops a theoretical foundation for this empirical approach using a specific-factors model of regional economies, clarifying how to construct measures of regional labor market exposure to trade liberalization and interpret the associated empirical results. In this setup, each region is a multi-industry specific-factors economy extending Jones (1975), in which labor is treated as mobile across industries and each location is endowed with a vector of industry-specific factors.[3] Examples of these regional industry-specific factors include natural resource inputs such as mineral deposits, land suitable for particular agricultural products, or industry-specific machinery. In the short run, both types of inputs are assumed to be immobile across regions. Because labor is assumed to be freely mobile across industries within region, wages are determined at the regional level, and one can derive a model-consistent estimating equation relating local wage changes to changes in tariffs across industries:

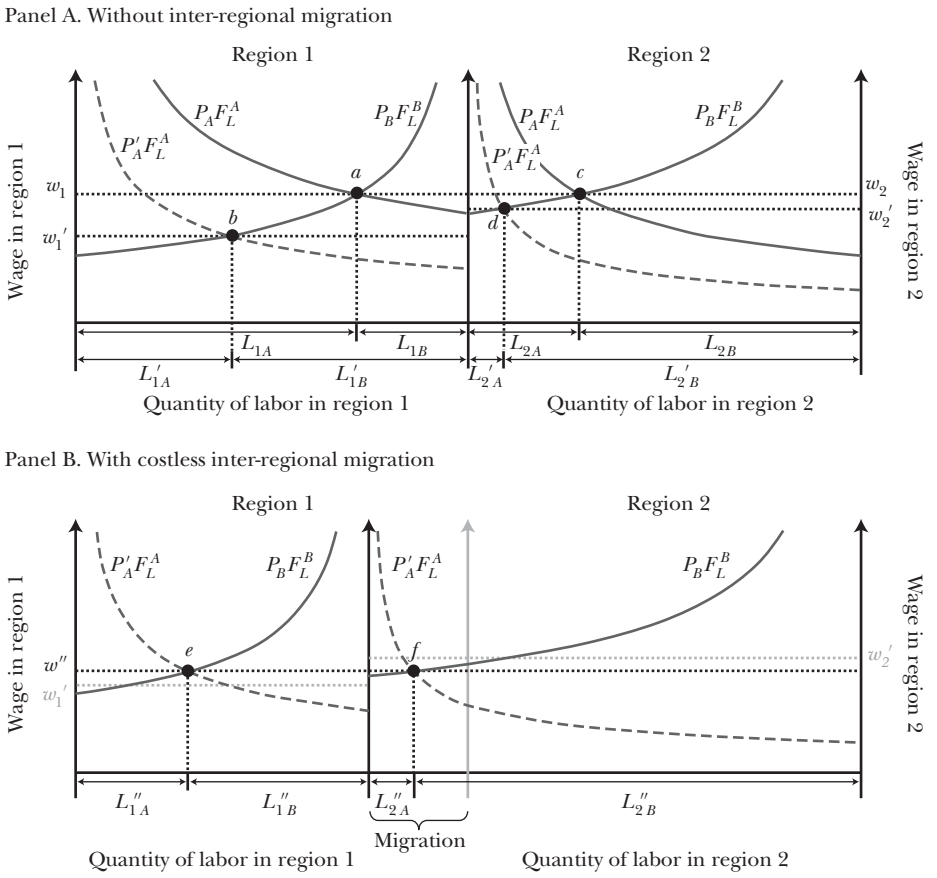
$$\Delta \ln w_r = \theta_0 + \theta_1 RTR_r + \varepsilon_r,$$

$$\text{where } RTR_r \equiv -\sum_{i \in T} \beta_{ri} \Delta \ln(1 + \tau_i) \quad \text{and} \quad \beta_{ri} \equiv \frac{\lambda_{ri} \frac{1}{\eta_i}}{\sum_{j \in T} \lambda_{rj} \frac{1}{\eta_j}}.$$

In words, the first line shows that the change in the log wage in region r, $\Delta \ln w_r$, relates to a weighted average of tariff changes, $\Delta \ln(1 + \tau_i)$, across tradable industries $i \in T$, which we refer to as the regional tariff reduction, or RTR_r. The second line defines the regional tariff reduction more specifically. The weights, β_{ri}, depend primarily upon the share of regional employment in each industry, λ_{ri}, but also upon the cost share of specific-factors in industry i, η_i (though in practice the latter term generally has little quantitative impact). As in any specific-factors model, when labor is the mobile factor, the effect of a given set of tariff changes on the real wage within a given region is ambiguous. Moreover, as in any difference-in-differences style regression, the regression equation compares outcome growth for units facing larger versus smaller shocks but cannot identify absolute effects on either group. Instead, it shows that the *relative* effect across regions is driven by a simple weighted average of tariff changes. This inability to identify aggregate effects in this type of cross-sectional regression parallels the same limitation in the earlier cross-industry analyses.

[3] Kovak (2013) additionally assumes that there are no trade costs across regions, so goods' prices are spatially equalized, that the liberalizing country is small, so domestic prices fall proportionately with one plus the tariff rate, and that nontradable goods' prices are determined in local equilibrium.

Figure 2
Specific-Factors Model of Local Labor Markets



Source: Authors' creation.
Note: In panel A each region is a specific-factors economy as in Figure 1, where labor L is the mobile factor. Production is Cobb-Douglas, with specific-factor cost share $\eta_i = 0.5$ in both industries, $L_r = 10$ in both regions, $T_{1A} = 1$, $T_{1B} = 0.4$, $T_{2A} = 0.4$, and $T_{2B} = 1$. Initially, $P_A = P_B = 1$, and after the price change, $P_A = 0.5$. Panel A shows the effect of the price reduction in changing wages w in each region. Panel B shows the effect of costless inter-regional migration in equalizing wages across locations. For visual clarity, the figure was generated assuming wages are initially equalized between the two regions, but this assumption is not necessary for any of the conclusions drawn here.

To provide intuition for the weighted-average structure of the regression equation, Figure 2 presents a graphical analysis in a context with two locations (regions $r = 1, 2$) and two industries ($i = A, B$). Each region is modeled as a specific-factors economy, as in Figure 1, and both regions face the same goods' prices. In panel A, labor L_r is mobile across industries within region r, and industry specific factors T_{ri} are immobile across both locations and industries. We assume that region 1 has a larger endowment of industry-A specific factor relative to industry B

than does region 2—that is, $T_{1A}/T_{1B} > T_{2A}/T_{2B}$. This difference in specific-factor endowments leads to differences in the marginal products of labor across regions such that, in equilibrium, the share of regional labor allocated to industry A is larger in region 1 than in region 2, as shown on the x-axes at the initial equilibrium points a and c.[4]

Starting from this initial equilibrium, assume the domestic price of good A falls due to a tariff reduction, lowering the value of marginal product for industry A in both regions. To restore regional equilibrium, workers shift out of industry A and into industry B until the values of marginal product (and thus wages) are equalized across the two industries. The magnitude of the resulting equilibrium wage decline will be greater when industry A is larger or if its labor demand is more elastic, because both of these features imply that a larger quantity of labor must leave the industry to drive a given decrease in the marginal product of labor. This reasoning explains the structure of the weights in the equation for regional tariff reduction, or RTR_r: larger industries have larger λ_{ri}, and with a Cobb-Douglas production function, the industry-specific labor demand elasticity is $1/\eta_i$. Therefore, in the resulting equilibrium, region 1's wage falls more than region 2's. Similarly, a change in P_B would have a larger effect on the wage in region 2 than in region 1, consistent with the weighted-average structure in the first equation above.

Kovak (2013) implements the regional research design implied by the specific-factors model in the context of Brazil's trade liberalization of the early 1990s, which involved large average tariff cuts and substantial variation in tariff cuts across industries. The empirical findings confirm that regional wages fall more (grow less) from 1991 to 2000 in regions facing larger tariff reductions. While this model predicts that $\theta_1 = -1$ in the absence of inter-regional migration, the empirical estimates are approximately -0.4. This difference suggests the presence of some amount of equalizing migration or random measurement error in the regional shock estimates. Nonetheless, as in Topalova (2007, 2010), the fact that such large regional effects remain nine years after the start of the liberalization episode implies the presence of important migration frictions across locations. In the following section, we place this effect estimate in the context of its evolution over time to learn more about the dynamic adjustment process.

While Topalova (2007, 2010) and Kovak (2013) examine countries lowering the barriers they impose on *imports*, McCaig (2011) studies the effects of lower barriers to *exports* in a dominant export destination. Specifically, he studies the effects of the US-Vietnam Bilateral Trade Agreement on local labor markets in Vietnam, finding larger poverty reductions in Vietnamese regions specialized in industries facing larger reductions in US tariffs, consistent with the notion that Vietnamese export goods facing larger US tariff declines experienced larger increases in relative prices. In a follow-up paper, McCaig and Pavcnik (2018) show that Vietnamese industries facing larger US tariff reductions experienced shifts into formal employment and

[4]For visual clarity, the figure was generated assuming wages are initially equalized between the two regions, but this assumption is not necessary for any of the conclusions drawn here.

increased productivity. In the context of the Canada-US Free Trade Agreement, Kovak and Morrow (2025) study the effects of tariff reductions on both imports and exports, finding adverse effects of Canadian tariff cuts and favorable effects of US tariff cuts for Canadian workers.

The literature on how trade policy can affect labor market outcomes has exploded in recent years, and a full literature review is beyond the scope of this paper, but a few additional threads are worth highlighting.[5] Workers may face frictions in switching industries and locations, meaning that industry-level and region-level shocks may be relevant for their outcomes. Hakobyan and McLaren (2016) examine the effect of NAFTA on US workers' wages, including terms capturing both an average regional shock, similar to RTR_r, and a direct industry-level shock. They find quantitatively and statistically significant effects in both dimensions, suggesting that Mexican import competition was relevant to US workers' wages based both on their location and industry of employment. More recently, Pierce, Schott, and Tello-Trillo (2024) use a similar approach to study the regional and industry effects of Chinese import competition in the United States, finding that regional effects were more important than industry effects.

While not studying a trade policy change per se, a prominent literature beginning with Autor, Dorn, and Hanson (2013) studies the local effects of increased imports from China in the US economy, using a weighted-average measure similar to the earlier formula for regional tariff reduction. They instrument for observed US imports from China using other countries' Chinese imports and find substantial and persistent relative reductions in manufacturing employment in regions facing larger growth in Chinese imports. This influential work has spawned many extensions and replications in other countries (for thorough literature reviews, see Autor, Dorn, and Hanson 2016; Autor et al. 2025). A complementary approach to studying the effects of Chinese import competition focuses on reductions in trade policy uncertainty associated with the US granting China Permanent Normal Trade Relations in 2001 (Handley and Limão 2017). While this change did not affect realized protection, it reduced the risk that the US Congress would raise tariffs on China, and that reduction in risk varied substantially across industries. Pierce and Schott (2016) show that US manufacturing industries facing larger reductions in tariff uncertainty experienced larger employment declines, and in Pierce and Schott (2020) they also use a regional weighted-average approach to document relative increases in unemployment rates and reductions in labor force participation in more exposed regions.

Most of the empirical literature discussed so far examines the effects of tariff reductions or increases in import competition, which induce relative declines in labor demand across regions or industries. A natural question is whether the effects are symmetric when tariffs rise and trade protection increases. Evidence from the 2018–2019 US-China trade war is beginning to shed light on this issue. Flaaen and

[5] In Dix-Carneiro and Kovak (2025), we provide a broader summary of the effects of globalization on inequality, including outcomes beyond the labor market, with a focus on Latin America.

Pierce (2024) show that, in the short run, industries receiving greater tariff protection experienced modest increases in employment, all else equal. However, these gains were offset by higher costs on imported intermediate inputs and by retaliatory tariffs imposed by China on US exports. On net, industries more exposed to tariff increases experienced relative declines in employment.

In an examination of a longer time horizon following the trade war, Autor, Beck, Dorn, and Hanson (2024) find that industries benefiting from tariff protection experienced increases in sales, but not in employment. Instead, higher sales reflected the ability of firms to raise prices and expand their capital stock. This evidence suggests that firms protected by tariffs responded both by exploiting reduced competitive pressure and by making production more capital intensive through automation and capital deepening. At the regional level, the authors find no statistically significant positive effects of tariff increases on employment. In contrast, retaliatory tariffs imposed by China generated relative employment declines in locations more exposed to these retaliatory measures, at least in the short run.

These findings offer three important lessons for the current debate over rising protectionism, particularly in large high-income economies. First, governments must be careful not to offset the intended benefits of tariff protection by raising the cost of imported intermediate inputs. Second, when a large country imposes tariffs on another large trading partner, retaliatory measures may substantially erode or even eliminate any employment gains from protection. Third, protecting industries may benefit firm owners without necessarily benefiting workers. Increased sales may reflect higher markups due to reduced competition and greater capital intensity in production, rather than higher labor demand.

Together, the industry- and region-level empirical approaches discussed in this section reveal a rich picture of the effects of trade policy on labor market outcomes, demonstrating the importance of labor market frictions even at decadal time scales. Yet all of the models and evidence discussed so far are inherently static, comparing outcomes in a single period before and a single period after a trade policy change. In the following section, we consider dynamic labor market responses to trade policy, highlighting recent evidence that in many cases went against conventional wisdom and required a rethinking of the simplest models.

Tariffs and Labor Market Dynamics: Empirics

A clear lesson from the static local-labor-market literature is that effects of tariff changes on local outcomes tend to be far more persistent than economists had initially expected. As just discussed, Topalova (2010) and Kovak (2013) document local impacts on poverty and wages that last for more than a decade, and Autor, Dorn, and Hanson (2013) find similarly durable effects of import competition from China on US unemployment and labor force participation. These findings were surprising at the time: the prevailing view was that inter-regional migration would be

strong enough to dissipate the effects of local labor-demand shocks on real wages and unemployment within a few years. That view was widespread in light of Blanchard and Katz (1992), who showed that US state-level unemployment and real wages tend to return to baseline over a horizon of roughly five to seven years, primarily through population outflows.

Given this background, Dix-Carneiro and Kovak (2017) fill an important gap in the literature by placing the snapshot provided by Kovak (2013) in the context of the long-run adjustment path. They document how regional employment and wages evolved year by year following liberalization, estimating dynamic responses of local labor market outcomes to region-specific tariff declines with this regression:

$$y_{r,t} - y_{r,0} = \theta_t RTR_r + \alpha_t + \epsilon_{r,t},$$

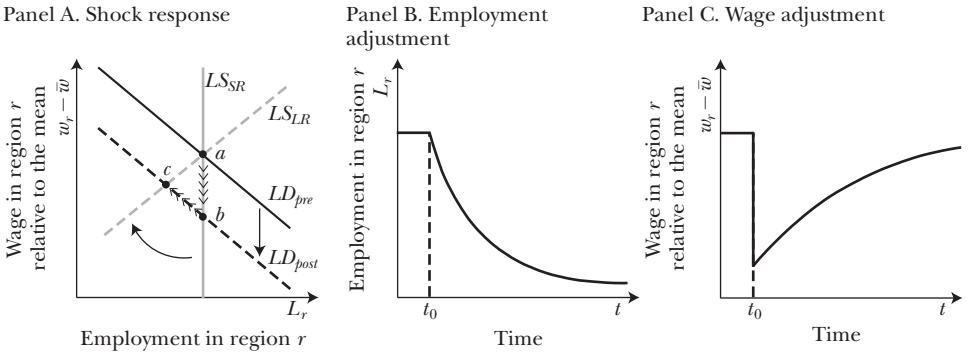
where the left-hand side is the change in region r's outcome y between period 0 (just before liberalization) and period t, RTR_r is the regional tariff reduction already discussed, α_t is a year fixed effect, and $\epsilon_{r,t}$ is a residual term. Importantly, the local effect of trade liberalization, θ_t, is allowed to vary freely over time. Because Brazil's liberalization was effectively a one-time shock concentrated between 1990 and 1995, the trajectory of θ_t directly traces short-, medium-, and long-run regional impacts of the fixed tariff shock captured by RTR_r.

In standard spatial models, a permanent decline in local labor demand—for example, a fall in labor demand induced by price declines faced by regional producers—leads to an immediate decline in wages, because workers are approximately immobile across regions in the short run and labor supply is effectively inelastic. As already shown in panel A of Figure 2, when a change in trade policy reduces the relative price of a region's most important products, it drives a wage gap between the more affected and less affected locations, with the magnitude of the gap reflecting the size of the labor demand shock. Panel A of Figure 3 illustrates the same effect in a standard supply-and-demand framework as the equilibrium moves from point *a* to point *b* in the short run.

The wage gap gives workers an incentive to move away from the negatively affected region to the rest of the country. In Figure 2, when a worker moves from region 1 to region 2, the *x*-axis shrinks by one unit in region 1 and grows by one unit in region 2, shifting the middle *y*-axis to the left along with the value of marginal product curves measured with respect to that axis ($P_B F_L^B$ in region 1 and $P_A F_L^A$ in region 2). As this migration progresses, the middle axis shifts further left, the equilibrium wage in region 1 rises, and the wage in region 2 falls. In the extreme case of perfect inter-regional mobility in the long run, this process continues until wages are equalized across locations, as shown in panel B of Figure 2.

The same migration process appears in Panel A in Figure 3 as a clockwise rotation of the labor supply curve; as workers gradually move away from the negatively affected region, labor supply becomes more elastic, and the equilibrium point slowly moves along the labor demand curve from *b* to *c*. Panel B shows the gradual but permanent decline in employment in the negatively affected region. When

Figure 3
A Model of Dynamic Labor Market Adjustment



Source: Authors' creation.
Note: Panel A shows the initial labor market equilibrium at point *a*, at the intersection of the pre-shock labor demand curve, LD_{pre}, and the short-run labor supply curve, LS_{SR}. After the negative tariff-induced labor demand shock, the labor demand curve moves down to LD_{post}, and the equilibrium moves to point *b* in the short run. Then, as workers migrate out of the negatively affected location, the labor supply curve gradually rotates clockwise to become the long-run labor supply curve, LS_{LR}. Panel B shows the predicted time path of regional employment, L_r, and panel C shows the time path of the regional wage relative to the cross-regional average, $w_r - \bar{w}$.

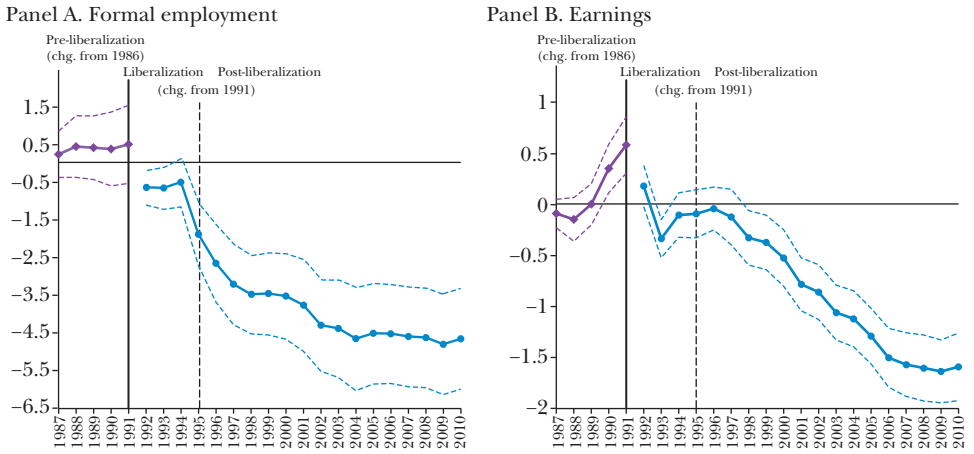
labor demand falls in the adversely affected region, wages drop on impact, but as workers gradually move away, wages partially recover. Whether wage gaps persist in the long run depends on the long-run labor supply elasticity—that is, on whether workers are perfectly or imperfectly mobile across regions in the long run. Panel C shows the resulting wage path: short-run effects are more negative, while long-run impacts are smaller in magnitude.

The response of Brazilian regional formal employment closely follows the predictions of this spatial model, as Dix-Carneiro and Kovak (2017) show. Panel A of Figure 4 plots the evolution of θ_t when the regression equation is estimated using the log of formal employment as the outcome *y*. Regions more exposed to tariff-induced import competition experience gradual relative declines in formal employment, with effects stabilizing roughly 15 years after the start of the liberalization episode. In sharp contrast, the behavior of log wages diverges markedly from the model's predictions. Rather than converging as workers migrate to equalize earnings across regions, the negative relative wage impacts deepen over time and stabilize only 17 to 20 years after liberalization begins, as shown in panel B of Figure 4.[6]

Two main mechanisms explain this behavior. First, capital adjusts slowly across regions. As capital flows out of regions more exposed to import competition, the initial drop in labor demand is gradually amplified: a shrinking capital stock lowers

[6]Note that due to a lack of hours information in most administrative data sources, including in Brazil, research in this area rarely distinguishes between monthly earnings and hourly wages.

Figure 4
Formal Employment and Earnings Dynamics

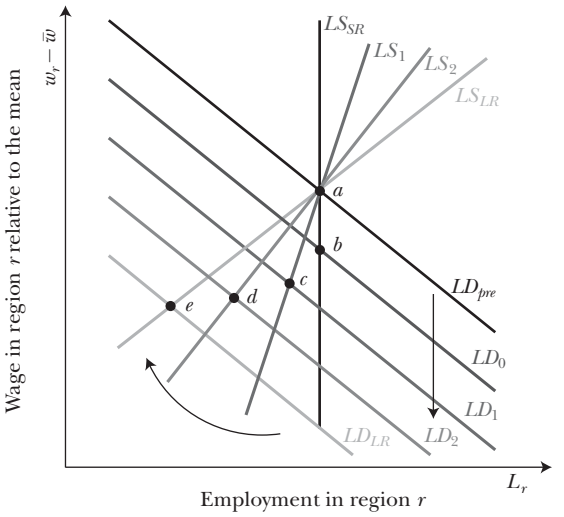


Source: Panel A replicates Figure 4 and panel B replicates Figure 3 of Dix-Carneiro and Kovak (2017).
Note: Each point reflects an individual regression coefficient, θ_t.

the marginal product of labor and pushes labor demand further down over time. Labor demand therefore falls on impact and keeps falling as capital relocates toward higher-return regions. Second, agglomeration economies further magnify the effects of these shocks; as employment declines, so does productivity, reinforcing the reduction in local labor demand. Together, these channels imply that labor-demand dynamics gradually amplify the initial effects of the liberalization shock, and wage impacts remain persistent because workers are imperfectly mobile across regions (that is, long-run labor supply is not perfectly elastic).

These mechanisms are illustrated in Figure 5. Following trade liberalization, labor demand in a region more exposed to tariff cuts falls on impact, shifting the equilibrium from point *a* to point *b*. Because short-run labor supply is inelastic, wages drop immediately. Over time, two forces unfold: (1) capital in the affected region depreciates and new investment flows elsewhere, pushing labor demand further down; and (2) workers gradually move out, rotating the labor supply curve clockwise. As these adjustments play out, the equilibrium shifts from point *b* to points *c*, *d*, and *e*. In this example, long-run wage effects exceed short-run effects, and employment losses are also amplified relative to those in Figure 3. Moreover, as labor flows out, it reduces local productivity—given the presence of agglomeration economies—which further reinforces the decline in labor demand caused by capital adjustment. The qualitative path for wages depends on which adjustment is faster: the downward shift in labor demand or the clockwise rotation of labor supply. The main takeaways from Dix-Carneiro and Kovak (2017) are that tariff-induced shocks generate meaningful regional dynamics and long adjustment paths. In Brazil's case, the economy required 17 to 20 years to adjust fully to the liberalization episode.

Figure 5
A Model of Dynamic Labor Market Adjustment with Dynamics in Labor Demand



Source: Authors' creation.
Note: As in Figure 3, the *y*-axis measures the regional wage relative to the cross-regional average, $w_r - \bar{w}$, and the *x*-axis measures regional employment, L_r. The initial labor market equilibrium is at point *a*. After the negative tariff-induced labor demand shock, the labor demand curve falls to LD_0 and the equilibrium point moves straight downward to point *b*. As capital reallocates away from the region and lost agglomeration externalities further reduce productivity, local labor demand falls over time to LD_1, LD_2, and LD_{LR}. At the same time, workers slowly migrate away, rotating the labor supply curve from LS_{SR} to LS_1, LS_2, and LS_{LR}. The equilibrium then moves through points *c*, *d*, and *e*.

Their findings also underscore frictions in workers' geographic mobility and show that agglomeration forces and the spatial reallocation of capital are central to understanding the observed adjustment process.[7]

In a follow-up paper, Dix-Carneiro and Kovak (2019) investigate how the same tariff-induced local shocks affected individual workers, focusing on the margins along which workers adjusted. Using administrative panel data, they show that workers initially employed in regions more exposed to import competition experienced formal-sector employment declines that grew over time. These workers were able to shift toward formal nontraded activities—especially low-paying jobs—but this adjustment margin was more than offset by larger losses in traded sectors. Using Brazilian decennial census data, the authors further document that more exposed regions saw increases in nonemployment in the medium run (1991–2000). In the long run (1991–2010) employment recovered, but largely through the expansion of informal work, which is widely viewed as lower-quality (Perry et al. 2007). Taken

[7]We encourage researchers to examine similar dynamics in other contexts with one-time trade policy shocks to assess the extent to which the results from Brazil generalize to other contexts.

together, these patterns indicate substantial disruption in local labor markets: firms exit in the short and medium run and new entry is limited, leaving displaced formal-sector workers unemployed for extended periods. Over time, however, local employment rebounds as workers shift into the informal sector, which appears to function as a buffer for trade-displaced workers. In its absence, the effects of negative labor-demand shocks on unemployment would likely have been larger and longer lasting—a hypothesis corroborated by Ponczek and Ulyssea (2022) and supported by the structural framework in Dix-Carneiro, Goldberg, Meghir, and Ulyssea (2026).

Tariffs and Labor Market Dynamics: Theoretical and Quantitative Insights

These empirical results, documenting complex and dynamic labor market responses to trade policy changes, have motivated the development of trade models incorporating salient features of labor markets, including search frictions and unemployment; worker mobility frictions across sectors, regions, and occupations; labor market regulations and informality; and transitional dynamics. These features allow the models to capture nuanced and realistic distributional impacts of trade policy, replicating key empirical patterns and allowing researchers to quantify the aggregate and distributional effects of trade along the entire transition path.

From Static to Dynamic: Introducing Mobility Frictions and Transitional Dynamics

Traditional models of trade capture either short-run effects (as in the specific factors model) or long-run outcomes (as in Heckscher-Ohlin or Ricardian models), but they offer no guidance on what happens during the transition—precisely the period over which policymakers worry most about the costs of trade liberalization. In particular, these models say little about the magnitude of the costs workers face as they adjust, how long the adjustment will take, or the lifetime impacts on workers across regions and industries once mobility frictions and transitional dynamics are taken into account. Such traditional models also cannot speak to anticipatory effects of policies announced in advance. These core questions in contemporary debates over trade policy simply cannot be addressed within static frameworks. The first generation of structural labor-trade models—including the framework of Artuç, Chaudhuri, and McLaren (2010)—filled this gap by introducing forward-looking workers and costly mobility across industries.

These dynamic models generate slow and gradual adjustment to trade shocks, as forward-looking workers face idiosyncratic and time-varying mobility costs. This structure generates aggregate adjustment paths that capture underlying mechanisms in which workers delay moving for various reasons: better opportunities arise at different times; many are initially reluctant to relocate; uncertainty about future prospects resolves only gradually; and workers in declining sectors often lack the skills needed elsewhere. Together, these forces explain why labor reallocation is slow even when long-run gains from trade are positive, and they help us characterize

the dynamic gains from trade: labor market frictions delay the realization of benefits from comparative advantage and specialization, lowering the present value of those gains.

From Representative Worker to Rich Heterogeneity

Still, even these newer dynamic models do not allow for differences in adjustment across workers: in practice, some individuals move quickly while others barely adjust, but the framework discussed above treats them as responding in the same way. Policymakers may want to identify which groups are more vulnerable to changes in trade policy and to design effective policies that compensate those who lose, thereby sustaining support for globalization. These distributional concerns—together with the evidence that adjustment varies widely across workers, sectors, occupations, and regions—motivated researchers to incorporate richer forms of worker heterogeneity.

Young and old workers, skilled and unskilled workers, and individuals with different occupational profiles respond very differently to trade-induced shifts in labor demand across sectors. Incorporating this heterogeneity—through overlapping generations, worker comparative advantage across sectors (observed or unobserved), or sector-specific experience—substantially changes both the quantification and the interpretation of mobility frictions. Dix-Carneiro (2014) shows that once selection and sorting of workers across sectors are taken into account, many workers face far more modest sectoral mobility costs than those estimated using the methodology of Artuç, Chaudhuri, and McLaren (2010), which assumed workers are identical. Moreover, because older workers tend to be less flexible while new entrants are more mobile, part of the aggregate sluggishness reflects generational turnover: the economy adjusts gradually as less adaptable cohorts retire and more adaptable cohorts enter. These differences in mobility frictions mean that tariff changes generate distributional consequences not only across sectors—as in the specific-factors model or Artuç, Chaudhuri, and McLaren (2010)—but also within sectors, among workers with different abilities to adjust.

In addition, Traiberman (2019) finds that adjustment can be slow in response to sectoral demand shocks not because changing *sectors* is particularly difficult, but because switching *occupations* is the main hurdle—and occupational composition varies systematically across sectors. This perspective underscores that workers' ability to adjust depends heavily on their initial tasks and their task-specific human capital. Because all of these frameworks embed worker heterogeneity directly, they can also be used to simulate compensation policies for the groups hurt by trade, including spillovers onto workers who do not receive support.

From Partial to Full General Equilibrium with Labor Market Frictions

The frameworks discussed so far are all partial-equilibrium in nature. They are built on small open-economy models in which changes in a country's trade policy do not affect world prices or global supply and demand. This prevents them from capturing how trade policy changes affect world prices, global patterns of

specialization, and the resulting gains from trade, while also limiting the counterfactuals they can analyze. For instance, these models cannot simulate how China's productivity growth reshaped global production and labor markets.

To move beyond these limitations, a third step in the literature—initiated by Caliendo, Dvorkin, and Parro (2019)—embeds labor market frictions into multi-country general-equilibrium models of trade built on the workhorse frameworks of Eaton and Kortum (2002) and Caliendo and Parro (2015). These Ricardian, multi-sector environments incorporate labor market dynamics and provide a unified framework for studying the gains from trade and its distributional effects when adjustment frictions matter, also accounting for how shocks in one country propagate to others. Moreover, they are able to estimate aggregate effects of changes in trade policy, overcoming an important limitation of the reduced-form approaches discussed above.[8]

Another key advantage of these general-equilibrium settings is that they describe full adjustment paths, not just steady states. This makes it possible to evaluate policies such as tariffs or study the effects of technological change in environments where reallocation is gradual and varies across regions and sectors. As a result, structural models of trade are now better aligned with the questions policymakers increasingly care about: not only how much trade affects welfare, but when, where, and for whom those effects arise.

How Macro Conditions and Institutions Shape Adjustment

These frameworks are also flexible enough to analyze how macroeconomic conditions and labor market institutions shape the labor market consequences of trade shocks. A central example is relaxing the balanced-trade assumption, which is common in traditional trade models. Under balanced trade, import shocks must be accompanied by simultaneous export expansions, allowing workers to move from import-competing sectors into growing exporting industries. But when import competition coincides with rising trade deficits, this mechanism breaks down. The economy absorbs the import shock without a corresponding export boom, making unemployment effects larger and more persistent than what balanced-trade models suggest—unless services or other less tradable sectors expand enough to absorb workers displaced by import competition.

In an examination of the importance of trade imbalances for the adjustment process, Dix-Carneiro, Pessoa, Reyes-Heroles, and Traiberman (2023) develop a quantitative general-equilibrium model of trade with frictional labor markets and aggregate saving and borrowing decisions that generate trade imbalances. This structure allows the model to capture episodes like China's "saving glut," in which shifts in foreign saving behavior drive persistent global imbalances independently

[8]A key motivation of Caliendo, Dvorkin, and Parro (2019) is to develop a general equilibrium model that rationalizes the relative, reduced-form effects documented by Autor, Dorn, and Hanson (2013) and to use this framework to infer the aggregate welfare and labor market effects of the China shock in the United States.

of productivity or changes in trade costs. A key insight from this and related work is that allowing trade imbalances to adjust in response to import shocks substantially changes the dynamics of labor market adjustment and unemployment. This work implies a significantly stronger role for developments in the Chinese economy in explaining the decline in US manufacturing employment than balanced-trade models do, because rising trade deficits amplify the reallocation of labor toward services and other less tradable sectors.

In a similar spirit, other models introduce institutional frictions such as downward wage rigidity to generate unemployment responses to negative shocks (Rodríguez-Clare, Ulate, and Vásquez 2026). Related work, building on the evidence in Dix-Carneiro and Kovak (2017), incorporates forward-looking capital accumulation and shows that slow capital adjustment interacts with labor mobility frictions, prolonging the local scarring effects of trade and delaying convergence (Kleinman, Liu, and Redding 2023).

A complementary line of research examines how institutional features—especially labor market regulations and the prevalence of informality—shape adjustment in developing economies. In environments where regulations are imperfectly enforced, firms choose between operating formally or informally, and this choice interacts with trade shocks in economically meaningful ways (Dix-Carneiro, Goldberg, Meghir, and Ulyssea 2026). Because more productive firms tend to operate formally while smaller and less productive firms remain informal, trade reforms reallocate resources across activities that face very different regulatory burdens. The central insight from this approach is that the gains from trade depend not only on comparative advantage, but also on how tariff changes shift activity between formal and informal production. When trade liberalization expands the more productive, but more heavily distorted, formal sector, overall gains from trade can be substantially larger than in otherwise similar economies without informality.[9]

Stringent labor market regulations such as firing costs and minimum wages can also significantly delay adjustment following trade liberalization (Kambourov 2009; Ruggieri 2022). Restrictive labor institutions slow the reallocation of workers across sectors after a trade reform, muting the subsequent gains in output, productivity, and welfare. The overarching message is that trade reforms are most effective when paired with labor market reforms that facilitate reallocation.

Summary

Taken together, the newly-developed structural frameworks give us a much richer view of how labor markets adjust to trade shocks. Dynamic models make it possible to study transitions, including how long adjustment to trade policy is likely

[9]This showcases the distinction between relative and aggregate effects. In general equilibrium, Dix-Carneiro, Goldberg, Meghir, and Ulyssea (2026) show that trade liberalization shifts activity toward the *formal* sector in the aggregate, while in partial equilibrium Dix-Carneiro and Kovak (2019) find that regions facing larger tariff reductions experience relative growth (or smaller contractions) in the *informal* sector. These results are reconciled if trade liberalization reduces aggregate informality, but regions more exposed to import competition experience slower declines in informal employment.

to take and who will bear the costs. Introducing worker heterogeneity reveals differences in mobility across workers, sectors, occupations, and regions, and allows for a more detailed characterization of distributional consequences of trade policy shifts. Embedding these ideas in multi-country general equilibrium demonstrates how shocks in one country transmit to others and more accurately quantifies the gains from trade when labor market frictions and slow adjustment matter. These advances have made quantitative models of trade and labor markets more realistic and therefore more useful for policymakers evaluating past trade reforms and forecasting the labor market consequences of future policy changes. Reflecting this progress, government agencies increasingly rely on these frameworks to evaluate not only past reforms but also prospective policies such as tariff changes.[10]

Smoothing Policies

Given the extensive evidence that trade shocks generate significant disruptions in labor markets, a natural question arises: How can governments mitigate the costs to workers and ensure that the gains from trade are shared more equitably? A voluminous body of literature studies the effects of active labor market policies targeting the broader population of displaced or searching workers, but there is much less evidence on policies specifically designed to compensate workers in occupations or sectors exposed to foreign competition.[11]

Before discussing policy options, it is important to understand three central arguments for instituting programs specifically targeting workers negatively affected by changes in trade policy. First, consistent with our earlier discussion, freer trade generally increases aggregate welfare while still leaving some groups of workers worse off. In this case, only by providing compensation to workers who lose from freer trade can it become Pareto-improving, making all agents in the economy weakly better off. Second, compensating those who lose from trade is essential for the successful and durable implementation of pro-trade policies. Because the benefits of trade are often diffuse and the costs concentrated, political opposition to a liberal trade regime is likely to arise absent compensation (Magee 2003). For example, Kim and Pelc (2021) find that US counties in which workers

[10] For example, in papers prepared for the US International Trade Commission, Riker (2022, 2024) adapt models from Caliendo, Dvorkin, and Parro (2019) and Dix-Carneiro, Pessoa, Reyes-Heroles, and Traiberman (2023) to examine the effects of productivity growth in China's Plastics and Rubber Products industry and of proposed labor provisions in trade agreements designed to strengthen workers' bargaining power. Likewise, in a paper prepared for the Brazilian Secretariat for Strategic Affairs, Góes et al. (2019) use the framework of Caliendo, Dvorkin, and Parro (2019) to assess a proposed tariff liberalization in Brazil, while World Bank (2025) employs the same framework to study the effects of trade liberalization and labor market reforms in South Asia.

[11] For surveys of the broader literature on the effects of active labor market policies, see Crépon and van den Berg (2016), Card, Kluve, and Weber (2018), and Bown and Freund (2019). These papers tend to find wide variation in the measured effectiveness of worker-support policies including job training, public or private employment subsidies, and job-search assistance.

receive approval for trade-related support from the Trade Adjustment Assistance (TAA) program exhibit reduced demands for protectionism in comparison to other counties facing similar import competition. Third, the demand for trade-displaced workers' occupational experience and skills may fall throughout the labor market as a result of trade liberalization such that they struggle to find reemployment even more than the average unemployed worker. If so, targeted support and training programs may be particularly helpful to trade-displaced workers in gaining new skills and transitioning into growing occupations.

These arguments motivated the creation of the US Trade Adjustment Assistance program, which operated in its modern form from 1974 to 2022 and was the largest program explicitly compensating workers who experienced job loss due to import competition or offshoring.[12] The program provided a variety of benefits to trade-displaced workers, with the most important provisions being extended unemployment insurance payments and subsidized job training. These benefits were designed both to compensate workers for their trade-induced job loss and to help them train for employment in another sector or occupation. Hyman (2018) studies the effects of TAA using the quasi-random assignment of eligibility petitions to examiners with different baseline probabilities of approving or denying eligibility. The program substantially increases earnings and employment for eligible workers, who are more likely to switch industries and to move to stronger labor markets. These results are encouraging, as they imply that with sufficiently generous benefits, trade-displaced workers' outcomes can be substantially improved in a cost-effective manner.

While the main provisions of the Trade Adjustment Assistance program subsidize trade-displaced workers while unemployed and in training, from 2002 onward the program included an alternative benefit known as "wage insurance," which provided an employment subsidy incentivizing speedy reemployment. Specifically, if a trade-displaced worker quickly found reemployment at a lower wage than in their predisplacement job, they could receive a subsidy covering up to half of the gap between their old and new wages for up to two years (subject to caps on cumulative benefit payments and on reemployment earnings). In theory, by making lower-paying jobs more attractive, this subsidy structure should shorten costly unemployment spells but could also lead to poor job matches. Hyman, Kovak, and Leive (2024) take advantage of the fact that the subsidy was available only to workers age 50 or over to identify its effects using a regression-discontinuity design, finding that the program reduced unemployment durations without meaningfully lowering wages or changing other job characteristics. It was also self-financing, with higher tax revenue from increased employment and savings from reduced benefits exceeding program costs. Wage insurance is therefore a promising option for

[12] While the program stopped considering new eligibility petitions on July 1, 2022, workers previously deemed eligible were able to continue receiving benefits after that date, consistent with their eligibility upon receiving certification. Note also that from 2002 onward, TAA covered workers indirectly affected by imports or offshoring, in industries upstream or downstream from directly affected industries.

helping trade-displaced workers quickly return to work, particularly among those for whom retraining may not be effective, feasible, or available.

An important caveat to the favorable findings for standard Trade Adjustment Assistance and wage insurance is that the programs are relatively small (TAA served an average of 24,000 workers per year from 2009 to 2022), so their existence is unlikely to induce spillovers through general equilibrium. If a larger share of workers were eligible for these policies, it is likely that eligible workers' favorable outcomes would come partly at the expense of ineligible workers or that firms would change their hiring strategies to capture a portion of the subsidies. We are unaware of evidence on these equilibrium effects in the specific context of policies targeting trade-displaced workers, but Crépon et al. (2013) document evidence for equilibrium spillovers associated with another active labor market policy.

Given the evidence discussed earlier that trade policy changes can have spatially concentrated effects, support for geographic relocation may help workers migrate to stronger markets and speed labor market transitions. For example, workers could apply for financial support to search for jobs outside their local labor market or governments could subsidize moving costs upon finding a job elsewhere. In the context of Bangladesh, Bryan, Chowdhury, and Mobarak (2014) found that a modest subsidy induced a significant share of rural workers to migrate to urban areas during the preharvest season, with positive effects on income and consumption and persistent effects via repeated migration in subsequent years.

Another potential means of facilitating worker transitions is to use algorithmic job recommendation systems to lower search costs for jobs that might be a good fit for trade-displaced workers, but of which they might be unaware. In this journal, Carranza and McKenzie (2024) observe that in many countries, the vast majority of workers search for jobs through friends and relatives. As these networks are often concentrated by location and industry, it is likely that trade policy changes will affect network members' prospects similarly. Job recommendation tools can broaden opportunities available to workers by presenting job openings in distant labor markets or in alternative industries or occupations whose skill demands match a searching worker's experience. The evidence on these systems is thus far mixed, with some increasing employment probabilities and/or earnings (Altmann, Glenny, Mahlstedt, and Sebald 2022; Belot et al. 2025; Belot, Kircher, and Muller 2026; Le Barbanchon, Hensvik, and Rathelot 2023), and others minimally affecting workers' outcomes (Bächli, Lalive, and Pellizzari 2025; Ben Dhia et al. 2022; Bied et al. 2026). An informal comparison of these papers suggests that effects are more likely to be favorable when searching workers are incentivized to engage meaningfully with the recommendation system and when job recommendations are based on observed worker transitions, though more evidence is needed.

Trade policy reforms themselves could also be adjusted to account for the costs associated with labor market frictions. For example, the government could plan a gradual change in policy to avoid abrupt labor market disruptions, allowing firms and workers time to prepare and adjust. The cost of this gradualism would be to delay the long-run gains from trade, with the potentially offsetting benefit of

reducing short-run adjustment costs. As discussed in the preceding section, trade reforms can also be coupled with reforms relaxing labor market regulations, with the goal of making labor markets more flexible and helping trade-displaced workers find reemployment more quickly (Kambourov 2009).

Finally, we conclude by emphasizing that, ideally, governments will implement policies for which there is evidence of success and that have undergone credible cost-benefit analysis. However, even imperfect policies may be beneficial if they compensate those who lose the most due to globalization and help maintain support for free trade. In the absence of some policy response supporting those experiencing the downsides of trade liberalization, the ongoing backlash against free trade is likely to continue, with trade wars reducing economic efficiency and eroding consumer purchasing power. Demonstrating concern for workers bearing the costs of globalization will likely be essential for political success in the years to come.

References

Altmann, Steffen, Anita M. Glenny, Robert Mahlstedt, and Alexander Sebald. 2022. "The Direct and Indirect Effects of Online Job Search Advice." IZA Discussion Paper 15830.

Artuç, Erhan, Shubham Chaudhuri, and John McLaren. 2010. "Trade Shocks and Labor Adjustment: A Structural Empirical Approach." *American Economic Review* 100 (3): 1008–45.

Attanasio, Orazio, Pinelopi K. Goldberg, and Nina Pavcnik. 2004. "Trade Reforms and Wage Inequality in Colombia." *Journal of Development Economics* 74 (2): 331–66.

Autor, David, Anne Beck, David Dorn, and Gordon H. Hanson. 2024. "Help for the Heartland? The Employment and Electoral Effects of the Trump Tariffs in the United States." NBER Working Paper 32082.

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

Autor, David H., David Dorn, and Gordon H. Hanson. 2016. "The China Shock: Learning from Labor-Market Adjustment to Large Changes in Trade." *Annual Review of Economics* 8: 205–40.

Autor, David, David Dorn, Gordon Hanson, Maggie R. Jones, and Bradley Setzler. 2025. "Places versus People: The Ins and Outs of Labor Market Adjustment to Globalization." In *Handbook of Labor Economics*, Vol. 6, edited by Christian Dustmann and Thomas Lemieux, 549–653. Elsevier.

Bächli, Mirjam, Rafael Lalive, and Michele Pellizzari. 2025. "Helping Jobseekers with Recommendations Based on Skill Profiles or Past Experience: Evidence from a Randomized Intervention." IZA Discussion Paper 17713.

Beaulieu, Eugene. 2002. "Factor or Industry Cleavages in Trade Policy? An Empirical Analysis of the Stolper–Samuelson Theorem." *Economics and Politics* 14 (2): 99–131.

Belot, Michèle, Bart K. de Koning, Didier Fouarge, Philipp Kircher, Paul Muller, and Sandra Phlippen. 2025. "Advising Job Seekers in Occupations with Poor Prospects: A Field Experiment." NBER Working Paper 33819.

Belot, Michèle, Philipp Kircher, and Paul Muller. 2026. "Do the Long-Term Unemployed Benefit from Automated Occupational Advice during Online Job Search?" *Economic Journal* 136 (673): 184–206.

Ben Dhia, Aïcha, Bruno Crépon, Esther Mbih, Louise Paul-Delvaux, Bertille Picard, and Vincent Pons. 2022. "Can a Website Bring Unemployment Down? Experimental Evidence from France." NBER Working Paper 29914.

Bied, Guillaume, Philippe Caillou, Bruno Crépon, Christophe Gaillac, Elia Pérennes, and Michèle Sebag. 2026. "A Job I Like or a Job I Can Get: Designing Job Recommender Systems Using Field Experiments." Preprint, arXiv. https://doi.org/10.48550/arXiv.2603.21699.

Blanchard, Olivier Jean, and Lawrence F. Katz. 1992. "Regional Evolutions." *Brookings Papers on Economic Activity* 23 (1): 1–61.

Bown, Chad P., and Caroline Freund. 2019. "Active Labor Market Policies: Lessons from Other Countries for the United States." Peterson Institute for International Economics Working Paper 19-2.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. 2014. "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh." *Econometrica* 82 (5): 1671–748.

Caliendo, Lorenzo and Fernando Parro. 2015. "Estimates of the Trade and Welfare Effects of NAFTA." *Review of Economic Studies* 82 (1): 1–44.

Caliendo, Lorenzo, Maximiliano Dvorkin, and Fernando Parro. 2019. "Trade and Labor Market Dynamics: General Equilibrium Analysis of the China Trade Shock." *Econometrica* 87 (3): 741–835.

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

Carranza, Eliana, and David McKenzie. 2024. "Job Training and Job Search Assistance Policies in Developing Countries." *Journal of Economic Perspectives* 38 (1): 221–44.

Colantone, Italo, Gianmarco Ottaviano, and Piero Stanig. 2022. "The Backlash of Globalization." In *Handbook of International Economics: International Trade*, Vol. 5, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 405–77. North-Holland.

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

Crépon, Bruno, and Gerard J. van den Berg. 2016. "Active Labor Market Policies." *Annual Review of Economics* 8: 521–46.

Deardorff, Alan V. 1994. "Overview of the Stolper-Samuelson Theorem." In *The Stolper-Samuelson Theorem: A Golden Jubilee*, edited by Alan V. Deardorff and Robert M. Stern, 7–34. University of Michigan.

Dix-Carneiro, Rafael. 2014. "Trade Liberalization and Labor Market Dynamics." *Econometrica* 82 (3): 825–85.

Dix-Carneiro, Rafael, Pinelopi Goldberg, Costas Meghir, and Gabriel Ulyssea. 2026. "Trade and Domestic Distortions: The Case of Informality." *Econometrica* 94 (2): 573–618.

Dix-Carneiro, Rafael, and Brian K. Kovak. 2017. "Trade Liberalization and Regional Dynamics." *American Economic Review* 107 (10): 2908–46.

Dix-Carneiro, Rafael, and Brian K. Kovak. 2019. "Margins of Labor Market Adjustment to Trade." *Journal of International Economics* 117: 125–42.

Dix-Carneiro, Rafael, and Brian K. Kovak. 2025. "Globalization and Inequality in Latin America." *Oxford Open Economics* 4 (S1): i376–99.

Dix-Carneiro, Rafael, João Paulo Pessoa, Ricardo Reyes-Heroles, and Sharon Traiberman. 2023. "Globalization, Trade Imbalances, and Labor Market Adjustment." *Quarterly Journal of Economics* 138 (2): 1109–71.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70 (5): 1741–79.

Ethier, Wilfred. 1974. "Some of the Theorems of International Trade with Many Goods and Factors." *Journal of International Economics* 4 (2): 199–206.

Flaaen, Aaron, and Justin Pierce. 2024. "Disentangling the Effects of the 2018-2019 Tariffs on a Globally Connected U.S. Manufacturing Sector." *Review of Economics and Statistics*. https://doi.org/10.1162/rest_a_01498.

Góes, Carlos, Alexandre Messa, Carlos Pio, Eduardo Leoni, and Luis Gustavo Montes. 2019. "Trade Liberalization and Active Labor Market Policies." In *Brazil: Boom, Bust, and the Road to Recovery*, edited by Antonio Spilimbergo and Krishna Srinivasan, 111–126. International Monetary Fund.

Goldberg, Pinelopi Koujianou and Nina Pavcnik. 2005. "Trade, Wages, and the Political Economy of Trade Protection: Evidence from the Colombian Trade Reforms." *Journal of International Economics* 66 (1): 75–105.

Goldberg, Pinelopi Koujianou, and Nina Pavcnik. 2007. "Distributional Effects of Globalization in Developing Countries." *Journal of Economic Literature* 45 (1): 39–82.

Goldberg, Pinelopi K., and Tristan Reed. 2023. "Is the Global Economy Deglobalizing? If So, Why? And What Is Next?" *Brookings Papers on Economic Activity* 54 (1): 347–96.

Hakobyan, Shushanik, and John McLaren. 2016. "Looking for Local Labor Market Effects of NAFTA." *Review of Economics and Statistics* 98 (4): 728–41.

Handley, Kyle, and Nuno Limão. 2017. "Policy Uncertainty, Trade, and Welfare: Theory and Evidence for China and the United States." *American Economic Review* 107 (9): 2731–83.

Hyman, Benjamin G. 2018. "Can Displaced Labor Be Retrained? Evidence from Quasi-Random Assignment to Trade Adjustment Assistance." In *Proceedings: Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*, Vol. 111, 1–70. National Tax Association.

Hyman, Benjamin G., Brian K. Kovak, and Adam Leive. 2024. "Wage Insurance for Displaced Workers." NBER Working Paper 32464.

Jones, Ronald W. 1965. "The Structure of Simple General Equilibrium Models." *Journal of Political Economy* 73 (6): 557–72.

Jones, Ronald W. 1971. "A Three-Factor Model in Theory, Trade, and History." In *Trade, Balance of Payments, and Growth: Papers in International Economics in Honor of Charles P. Kindleberger*, edited by Jagdish N. Bhagwati, Ronald W. Jones, Robert A. Mundell, and Jaroslav Vanek, 3–21. North-Holland.

Jones, Ronald W. 1975. "Income Distribution and Effective Protection in a Multicommodity Trade Model." *Journal of Economic Theory* 11 (1): 1–15.

Jones, Ronald W., and José A. Scheinkman. 1977. "The Relevance of the Two-Sector Production Model in Trade Theory." *Journal of Political Economy* 85 (5): 909–35.

Kambourov, Gueorgui. 2009. "Labour Market Regulations and the Sectoral Reallocation of Workers: The Case of Trade Reforms." *Review of Economic Studies* 76 (4): 1321–58.

Kemp, Murray C. and Henry Y. Wan. 1976. "Relatively Simple Generalizations of the Stolper-Samuelson and Samuelson-Rybczynski Theorems." In *Three Topics in the Theory of International Trade: Distribution, Welfare, and Uncertainty*, 49–59. North Holland.

Kim, Sung Eun, and Krzysztof J. Pelc. 2021. "The Politics of Trade Adjustment versus Trade Protection." *Comparative Political Studies* 54 (13): 2354–81.

Kleinman, Benny, Ernest Liu, and Stephen J. Redding. 2023. "Dynamic Spatial General Equilibrium." *Econometrica* 91 (2): 385–424.

Kovak, Brian K. 2013. "Regional Effects of Trade Reform: What Is the Correct Measure of Liberalization?" *American Economic Review* 103 (5): 1960–76.

Kovak, Brian K., and Peter M. Morrow. 2025. "The Long-Run Labour Market Effects of the Canada-U.S. Free Trade Agreement." *Review of Economic Studies* 92 (6): 4026–58.

Le Barbanchon, Thomas, Lena Hensvik, and Roland Rathelot. 2023. "How Can AI Improve Search and Matching? Evidence from 59 Million Personalized Job Recommendations." Preprint, SSRN. https://ssrn.com/abstract=4604814.

Magee, Christopher. 2003. "Endogenous Tariffs and Trade Adjustment Assistance." *Journal of International Economics* 60 (1): 203–22.

Magee, Stephen P. 1980. "Three Simple Tests of the Stolper-Samuelson Theorem." In *Issues in International Economics*, edited by Peter Oppenheimer, 138–53. Oriel Press.

Mayda, Anna Maria, and Dani Rodrik. 2005. "Why Are Some People (and Countries) More Protectionist than Others?" *European Economic Review* 49 (6): 1393–430.

Mayer, Wolfgang. 1974. "Short-Run and Long-Run Equilibrium for a Small Open Economy." *Journal of Political Economy* 82 (5): 955–67.

McCaig, Brian. 2011. "Exporting Out of Poverty: Provincial Poverty in Vietnam and U.S. Market Access." *Journal of International Economics* 85 (1): 102–13.

McCaig, Brian, and Nina Pavcnik. 2018. "Export Markets and Labor Allocation in a Low-Income Country." *American Economic Review* 108 (7): 1899–941.

Mussa, Michael. 1974. "Tariffs and the Distribution of Income: The Importance of Factor Specificity, Substitutability, and Intensity in the Short and Long Run." *Journal of Political Economy* 82 (6): 1191–203.

Perry, Guillermo E., William F. Maloney, Omar S. Arias, Pablo Fajnzylber, Andrew D. Mason, and Jaime Saavedra-Chanduvi. 2007. *Informality: Exit and Exclusion*. World Bank.

Pierce, Justin R., and Peter K. Schott. 2016. "The Surprisingly Swift Decline of US Manufacturing Employment." *American Economic Review* 106 (7): 1632–62.

Pierce, Justin R., and Peter K. Schott. 2020. "Trade Liberalization and Mortality: Evidence from US Counties." *American Economic Review: Insights* 2 (1): 47–64.

Pierce, Justin R., Peter K. Schott, and Cristina Tello-Trillo. 2024. "To Find Relative Earnings Gains after the China Shock, Look Outside Manufacturing and Upstream." NBER Working Papers 32438.

Ponczek, Vladimir, and Gabriel Ulyssea. 2022. "Enforcement of Labour Regulation and the Labour Market Effects of Trade: Evidence from Brazil." *Economic Journal* 132 (641): 361–90.

Revenga, Ana L. 1992. "Exporting Jobs? The Impact of Import Competition on Employment and Wages in U.S. Manufacturing." *Quarterly Journal of Economics* 107 (1): 255–84.

Revenga, Ana. 1997. "Employment and Wage Effects of Trade Liberalization: The Case of Mexican Manufacturing." *Journal of Labor Economics* 15 (S3): S20–43.

Riker, David. 2022. "Imports of Plastic and Rubber Products from China: An Application of the Caliendo Dvorkin Parro Model." US International Trade Commission Economics Working Paper 2022–05–D.

Riker, David. 2024. "Estimating the Economic Effects of Labor Provisions Using a Dynamic GE Model." US International Trade Commission Economics Working Paper 2024–04–C.

Rodríguez-Clare, Andrés, Mauricio Ulate, and José P. Vásquez. 2026. "Trade with Nominal Rigidities: Understanding the Unemployment and Welfare Effects of the China Shock." *Journal of Political Economy* 134 (2): 626–64.

Ruggieri, Alessandro. 2022. "Trade and Labor Market Institutions: A Tale of Two Liberalizations." Unpublished.

Samuelson, Paul A. 1987. "Joint Authorship in Science: Serendipity with Wolfgang Stolper." *Journal of Institutional and Theoretical Economics(JITE)/Zeitschrift für die gesamte Staatswissenschaft* 143 (2): 235–43.

Stolper, Wolfgang F., and Paul A. Samuelson. 1941. "Protection and Real Wages." *Review of Economic Studies* 9 (1): 58–73.

Topalova, Petia. 2007. "Trade Liberalization, Poverty and Inequality: Evidence from Indian Districts." In *Globalization and Poverty*, edited by Ann Harrison, 291–332. University of Chicago Press.

Topalova, Petia. 2010. "Factor Immobility and Regional Impacts of Trade Liberalization: Evidence on Poverty from India." *American Economic Journal: Applied Economics* 2 (4): 1–41.

Traiberman, Sharon. 2019. "Occupations and Import Competition: Evidence from Denmark." *American Economic Review* 109 (12): 4260–301.

World Bank. 2025. *South Asia Development Update, October 2025: Jobs, AI, and Trade*. World Bank.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 101–122

Should We Tax Trade? A Pigouvian Perspective

Arnaud Costinot and Iván Werning

The scene happens at Rochester, New York, in the 1970s. A soon-to-be-famous trade economist in his early 30s meets an already famous macroeconomist in his late 60s. The trade economist is visiting for the semester, teaching a new undergraduate class. The macroeconomist is there for the day and asks what the class is about. "It is about trade policy," the trade economist replies. "What is there to teach?" the macroeconomist quips. "Don't we all know free trade is optimal?"[1]

The goal of this article is to discuss why free trade may not always be optimal and, in such circumstances, what the optimal trade policy may be. We approach the problem of optimal trade policy as one of optimal technology regulation, with international trade as the technology of interest. Our focus is primarily normative, without apology. We will not try to explain why politicians set tariffs the way they do. Instead we will illustrate through a series of examples how theory and data can be combined in order to make progress on what trade policy ought to be.

To set the stage for our main analysis, we first offer a brief reminder of mercantilist ideas and the classical case for free trade. We then turn to the Theory of the Second Best, as a rationale for trade policy intervention, and the Targeting Principle, as an alternative defense of free trade. We conclude by discussing the limits of this defense and what optimal trade policy might entail in practice. Some of

■ *Arnaud Costinot and Iván Werning are Professors of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Their emails are costinot@mit.edu and iwerning@mit.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20251491.

[1] The two participants in the Rochester scene are Elhanan Helpman and Milton Friedman. We thank Elhanan Helpman for sharing this anecdote.

these themes are as old as the field of international economics itself. The novel and unifying perspective put forward in this article is a Pigouvian one.

The Pigouvian logic is familiar in environmental economics. If firms or consumers do not internalize the marginal impact of an economic decision on social welfare, such as pollution, then a Pigouvian tax should make them pay for it. As we will see, the same logic, broadly interpreted, is equally useful in international economics. Applied to the decision to trade with the rest of the world, it provides an intuitive way to characterize the structure of optimal trade policy, regardless of whether departures from free trade reflect concerns for efficiency or redistribution.

The core of this article is a series of formulas for optimal trade taxes and subsidies, as the case may be. Each formula highlights relevant sufficient statistics. For given values of these statistics, a formula provides conditions that trade taxes must satisfy in order to be optimal. If the import tariff currently imposed on a given good is lower than the one implied by the formula, then a tax reform that reduces imports of this good must be welfare-improving. If the observed tariff is higher instead, then raising imports from their current levels must raise social welfare, at least locally.

A key strength of our Pigouvian perspective, beyond its ability to accommodate various normative considerations, is its robustness to positive considerations. The specifics of the economic environment may affect how changes in trade policy map into changes in imports via income effects, economy-wide changes in factor prices, or input-output linkages. These general-equilibrium complications, however, have no bearing on the set of sufficient statistics that enter these formulas.

Another strength of our Pigouvian perspective is that it applies both when tariffs are the only policy instrument available and when they are not. Pigou's logic shifts the discussion over tariff policy away from absolutist claims that free trade is always optimal and also away from indiscriminate uses of tariff policy to address a wide variety of policy goals. Tariffs can be a targeted way to influence a foreign country for geopolitical or environmental reasons, provided that they affect international prices. Tariffs may also complement other imperfect domestic redistributive instruments, though existing empirical evidence suggests that the resulting optimal tariffs may be quantitatively small in this case.[2]

For expositional purposes, we often refer to import tariffs rather than trade taxes more broadly. The formulas below, however, apply without further qualifications to the case of import subsidies, export subsidies, and export taxes. All trade taxes are wedges between domestic and international prices, which the formulas

[2]Our analysis builds on classical discussions of the rationale for trade policy in Meade (1955), Johnson (1965), Bhagwati, Jones, Mundell, and Vanek (1971), and Corden (1974). Related discussions of the Targeting Principle and the Theory of the Second Best, applied to other policy instruments, can be found in Tinbergen (1952) and Lipsey and Lancaster (1956). Irwin (1996) provides an intellectual history of the case for free trade, which we have used to organize our theoretical analysis. Modern treatments of these ideas are developed in Dixit (1985) and Helpman and Krugman (1989). Some of the general themes developed in this article are also discussed in Costinot and Rodríguez-Clare (2025), but without the Pigouvian emphasis. Detailed derivations of the tariff formulas presented here can be found in Costinot and Werning (2023) and Adão, Becko, Costinot, and Donaldson (2025, 2026).

below characterize. Import tariffs and export subsidies are positive wedges, whereas import subsidies and export taxes are negative wedges.

Act I: The Mercantilist Fallacy

Mercantilists view international trade as a zero-sum game in which exports are good and imports are bad. Accordingly, a trade surplus reveals a country's success on the world markets, whereas a trade deficit indicates its failure. The mercantilist view has a long history. Irwin (1996) offers this classical statement from Mun (1664): "The ordinary means therefore to increase our wealth and treasure is by foreign trade, wherein we must observe this rule; to sell more to strangers yearly than we consume theirs in value." The same way firms ought to maximize their sales net of their expenditures, countries ought to maximize the value of their exports net of their imports.

In such a world, it might seem optimal for countries to impose import tariffs as a way to reduce their trade deficits. In practice, a desire to correct trade imbalances was one of the main drivers of the rise in protectionism during the 1930s, what Irwin (2012) refers to as the great trade policy disaster. Almost 100 years later, trade deficits remain salient in today's discussions about trade policy. The formula used to generate the "Liberation Day" tariffs announced by President Trump on April 2, 2025 is a recent manifestation of these ideas.

On the positive side, there is a question about whether tariffs can actually affect trade deficits. There is no doubt that tariffs can lower imports. But it is far from clear why the same import restrictions would leave exports unchanged. In fact, one of the most robust insights about trade policy is the equivalence between import tariffs and export tariffs due to Lerner (1936). From a macroeconomic perspective, it is not immediately obvious how tariffs would change a country's incentives to save and invest. But if tariffs do not affect national saving and domestic investment, how can they affect trade deficits?

In Costinot and Werning (2025), we have highlighted systematic reasons why permanently raising tariffs can lower trade deficits. Intuitively, even if a tariff is permanent, its incidence may vary over time. In periods of trade deficits, consumption and, in turn, imports tend to be higher. As a result, a permanent tariff has a disproportionate impact on the cost of living in these periods. All else equal, this provides incentives for consumers to spend less in periods of deficits, thereby reducing global imbalances.[3] The Liberation Day tariffs of April 2025 have generated a new round of predictions about the quantitative impact of tariffs on deficits (for example, Auclert, Rognlie, and Straub 2025; Caliendo, Kortum, and Parro

[3]This insight relates to Obstfeld and Rogoff's (2000) influential observation that trade costs between countries may limit the magnitude of global imbalances, as subsequently investigated by Eaton, Kortum, and Neiman (2016) and Reyes-Heroles (2016).

2025), but it is fair to say that despite recent empirical work, there remains considerable uncertainty about the impact of tariffs on trade deficits.

The fundamental issue with the mercantilist's perspective, however, is not positive, but normative. That is, why would the maximization of trade surpluses ever emerge as the relevant social welfare objective of a country? Imagine a radical government that chooses to prohibit the consumption of all goods in its economy. By doing so, it succeeds in turning a trade deficit into a trade surplus. Would anyone argue that social welfare must have increased as a result? Trade imbalances simply cannot be taken as a starting point for the welfare evaluation of trade policy. As repeatedly argued by trade economists, from Ricardo (1817) to Krugman (1996), countries are not like firms, and international trade is not a zero-sum game.

Act II: The Classical Case for Free Trade

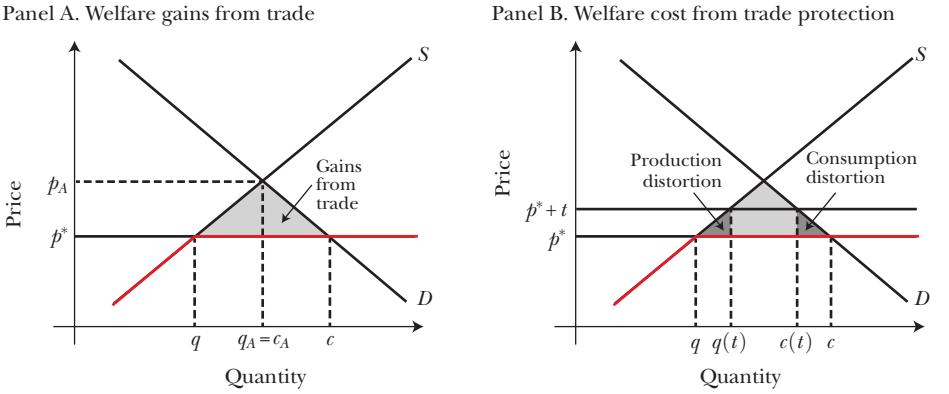
The classical case for free trade turns the mercantilistic logic on its head. It puts people rather than firms at the center of the analysis, with imports the ultimate benefit from international trade and exports its unfortunate cost. At its core, the classical case for free trade can be broken down into two parts. First, opening up to international trade is a form of technological progress. Second, technological progress is welfare-improving and should not be restricted.

That the ability to trade with the rest of the world is, ultimately, a form of technological progress is one of the most fundamental observations in international economics. Figure 1 offers a graphical representation of this idea. Consider a good, say electric vehicles, that can be either produced domestically or imported from abroad at some fixed price p^*. Another so-called outside good is used as numeraire and implicitly exported in the background. The supply curve S represents the marginal cost of producing electric vehicles domestically. The demand curve D represents the marginal benefit of consuming electric vehicles.

Under autarky, a country must produce what it consumes, $q_A = c_A$, leading to an autarky price p_A. Under free trade, it also has the option of importing from abroad. Hence the relevant "technology" is the lower envelope of the domestic supply curve S and the horizontal line going through p^*, which is the foreign export supply curve. Going from S to the lower envelope of S and p^* is technological progress.

One may be tempted to view the fact that imports are equal to domestic consumption minus domestic production, $m = c - q$, as an accounting identity that defines imports. Figure 1 suggests instead interpreting $c = q + m$ as a structural relationship that reflects a choice over two distinct technologies, domestic production (q) and imports (m), that can both be used to deliver consumption (c). If the new technology "imports" is deployed efficiently, as described in Figure 1, panel A, with domestic production occurring up to the point at which the domestic marginal cost is equal to p^* and domestic consumption occurring up to the point at

Figure 1
The Classical Case for Free Trade



Source: Provided by authors.
Note: Panel A describes the welfare gains of moving from autarky to free trade. Panel B describes the welfare costs of imposing an import tariff.

which the domestic marginal benefit is equal to p^*, then opening up to trade must raise welfare.

By how much? The total gains from trade can be measured by the area of the gray triangle. The gains reflect the fact that in the absence of international trade, some electric vehicles would still be consumed, but they would be produced at an unnecessarily high cost—the area of the triangle with base $q_A - q$. Furthermore, some electric vehicles would no longer be consumed, despite the fact that its marginal benefit is greater than its marginal cost p^*—the area of the triangle with base $c - c_A$. The sum of the two is the area under the import demand curve.[4]

A tariff is not quite as bad as shutting down all trade, but it lowers welfare by distorting the way the new technology of "imports" is being deployed. If a specific tariff t is imposed on imports of electric vehicles, it creates a wedge between the international price p^* received by foreign exporters and the domestic price p paid by domestic consumers and, in turn, received by domestic firms, with $p = p^* + t$.[5] This leads domestic firms to produce electric vehicles that should have been imported instead (a production distortion, corresponding to the area of the dark

[4]The fact that the welfare gains from trade are equal to the area under the import demand curve, properly defined, does not depend on the partial equilibrium nature of the economy depicted in Figure 1, as further discussed in this journal in Costinot and Rodríguez-Clare (2018).

[5]For expositional purposes, we describe all trade taxes in this article in specific rather than ad-valorem terms. A specific tariff is the dollar amount that an importer must pay in order to import one unit, whereas an ad-valorem tariff is charged as a percentage of the import price. It is important to note, however, that if t is the optimal specific tariff, then the optimal ad-valorem tariff is simply $t^{av} = t/p^*$ and vice versa. The equivalence between the two instruments reflects the fact that what matters is the wedge between p and p^*, not how it is implemented.

gray triangle with base $q(t) - q$) and consumers not to consume electric vehicles that should have been imported as well (a consumption distortion, corresponding to the area of the dark gray triangle with base $c - c(t)$). The welfare cost of trade protection is given by the sum of the two dark gray triangles in Figure 1, panel B.

Act III: The Limits of the Classical Case

If trade is a form of technological progress, it may be tempting to conclude that tariffs and other trade restrictions are necessarily a bad idea. There is no reason to throw rocks in your harbor, even if other countries previously did so, as Beveridge (1931) observed in response to the rise of protectionism in the 1930s. Notwithstanding the rhetorical appeal of the rocks-in-the-harbor metaphor, the classical case for free trade is only a profound insight, not a universal truth about the world. Indeed, there are potential reasons why one may want, under certain circumstances, to limit technological progress, from the development of robots to artificial intelligence. The same general reasons apply to international trade: namely, efficiency and redistribution.

For now, we assume that trade taxes are the only policy instruments, besides a uniform lump-sum transfer used to rebate tax revenue. To streamline our optimal tariff formulas, we consider separately the cases for trade policy as a tool for efficiency, domestic redistribution, and international redistribution. Combining these three cases is straightforward. The optimal trade tax when all three motives are present is simply the sum of the three terms appearing on the right-hand side of each formula below.

Trade Policy as a Tool for Efficiency

Key to the classical case—in favor of free trade and against trade protection—is the idea that the invisible hand of the markets always leads the economy towards efficiency. In such circumstances, expanding technological possibilities, be it via new scientific discoveries or simply cheaper imports, must raise welfare. More options cannot make you worse off if you always pick the best one.

There are well-known reasons, however, why market forces may fail to deliver efficiency. Textbook examples include externalities, imperfect competition, and nominal rigidities. Might some degree of trade policy intervention be optimal in this case? The Theory of the Second Best answers in the affirmative. If there are pre-existing distortions, then using even an imperfect policy instrument that creates its own distortions can lead to efficiency gains.

For concreteness, let us start with an economy subject to an externality e. Firms are perfectly competitive and there is a representative agent, so that redistribution is not relevant in the model and correcting the externality is the only motive for trade policy. The externality e may be domestic in nature, like air pollution or local knowledge spillovers, or international, like global carbon emissions or some geopolitical action. What matters for our purposes is that the externality creates a

wedge between the (net) social and private marginal costs of e, which we will simply refer to as social harm.

Next, consider the following hypothetical scenario. Starting from observed trade policy levels, the government engineers a small policy change designed to lower the imports of a given good g by $\Delta m_g < 0$. In the absence of general-equilibrium effects, the required policy change would be a small increase $\Delta t_g > 0$ in the tariff imposed on that good. In the presence of general-equilibrium effects, this may require adjusting other tariff lines to make sure that, out of all the goods subject to trade policy, only the imports of good g vary. General-equilibrium considerations would also affect how changes in trade policy map into changes in the externality e. Fortunately, none of these complicated considerations affect the basic structure of what the optimal trade tax on good g ought to be. A necessary condition for the observed tax t_g to be optimal is

$$\text{(Externality)} \qquad t_g = \textit{Social Harm}(e) \times \frac{\Delta e}{\Delta m_g},$$

with the social harm of raising the externality e equal to

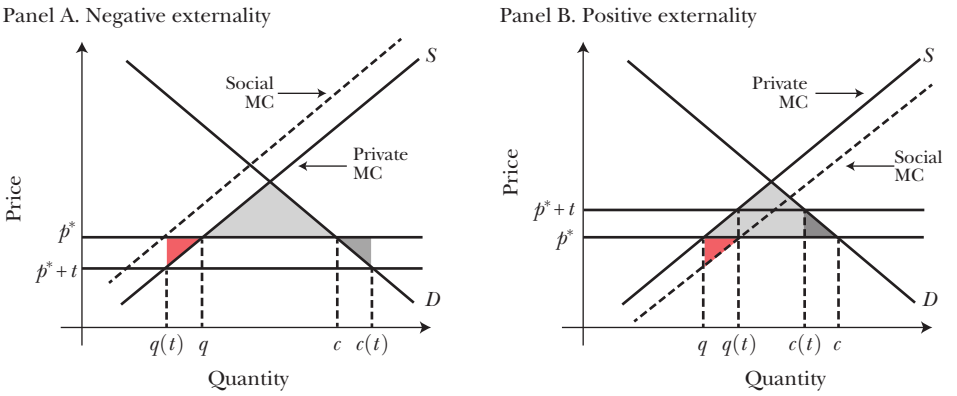
$$\textit{Social Harm}(e) = \textit{Social Marginal Cost}(e) - \textit{Private Marginal Cost}(e),$$

which could be positive or negative depending on the sign of the externality.

The tariff in the externality formula equates the marginal benefit of the policy reform in terms of fiscal revenue, $t_g \Delta m_g$, with its marginal cost in terms of efficiency, $\textit{Social Harm}(e) \times \Delta e$. Figure 2 illustrates the situation where the externality is caused by domestic production, so that $e = q$. If the externality is negative ($\textit{Social Harm}(e) > 0$) and imports of good g reduce externalities ($\Delta e / \Delta m_g < 0$), then the optimal tariff is a subsidy, $t_g < 0$, as described in Figure 2, panel A. The converse is true if the externality is positive, as described in Figure 2, panel B. In each case, the gains from correcting the externality are given by the area of the red triangles.

The standard Pigouvian logic is that if consumers or firms do not internalize the social harm of their own actions, then a tax should make them pay for it. When the action of interest is directly generating the externality e, such as "emitting carbon," the policy recommendation is to set a carbon tax equal to its social harm. When the action of interest m_g is "importing Brazilian beef" or "importing Indonesian palm oil," the logic is the same, but it requires adjusting the social harm of carbon by the impact of import restrictions on global carbon emissions, as summarized by the term $\Delta e / \Delta m_g$ in the externality formula. This term captures all the general-equilibrium considerations that affect the relationship between trade policy and emissions, including the possibility of carbon leakage across countries. Existing work on trade policy and climate change (for example, Kortum and Weisbach 2021; Hsiao 2026; Farrokhi and Lashkaripour 2025; Garci-Lembergman, Ramondo, Rodríguez-Clare, and Shapiro 2025) can be viewed as attempts to combine theory and data in order to provide empirically credible estimates of $\Delta e / \Delta m_g$. With these estimates at hand, along with an estimate of the social harm of carbon, commonly referred to as the

Figure 2
Trade Policy as a Tool for Efficiency



Source: Provided by authors.
Note: Panel A describes the welfare gains of imposing an import subsidy in the presence of a negative production externality. Panel B describes the welfare gains of imposing an import tariff in the presence of a positive production externality.

"social cost of carbon," one can then use the externality formula to evaluate the optimality of current trade policy for any given good.

The same logic applies to any externality, not just environmental ones. Historically, the infant industry argument for (temporary) tariffs is a classical departure from free trade that arises from such considerations, as discussed in Bardhan (1970). If there are positive externalities due to learning-by-doing in some industry—Irwin and Klenow (1994), Goldberg et al. (2024), and Barwick, Kwon, Li, and Zahur (2025) provide estimates of such learning effects for the semiconductor and electric vehicle industries—then the above formula suggests subsidizing imports of goods g that raise output in this industry, like intermediate inputs used in the industry, and taxing import of goods that reduce it. The same logic also applies in the case of geopolitical considerations. For instance, a country may care about national security and the military action e taken by another, as emphasized in recent work by Becko and O'Connor (2025), Becko, Grossman, and Helpman (2025), and Clayton, Maggiori, and Schreger (2026). If changes in imports can influence that decision, then Pigou's logic suggests restricting the imports of goods in proportion to their ability to influence this action, as again reflected in $\Delta e/\Delta m_g$.[6]

It should also be clear that there is nothing specific about externalities in the previous discussion. The same formula holds independently of why social and private incentives diverge. Suppose that because of imperfect competition, firms

[6]Compared to Becko and O'Connor (2025), Becko, Grossman, and Helpman (2025), and Clayton, Maggiori, and Schreger (2026), we abstract from the issue of how trade policy may also be used as a threat off-the-equilibrium path.

charge different markups on the goods that they sell, leading to misallocation in the economy, a classical theme in Helpman and Krugman (1989). As a result, the value of the marginal product of the factors that firms employ, say labor, is not equalized across these: employment in high-markup firms, whose goods are undersupplied, has higher marginal value than in low-markup firms, whose goods are oversupplied. Let e_n denote employment in a given firm n. In this situation, the wedge between the (net) social and private marginal costs of e_n is the opposite of firm n's markup τ_n. Pigou's logic then implies setting a tariff on the imports of good g equal to its damage on misallocation in the economy,

$$\text{(Misallocation)} \qquad t_g = -\sum_n \tau_n \times \frac{\Delta e_n}{\Delta m_g}.$$

This misallocation formula holds true regardless of whether firms are monopolistically competitive or whether they are strategic and compete à la Bertrand, Cournot, or some variation. This is also true when the difference between private and social marginal cost τ_n arises from taxes faced by firms, in which case the "Misallocation" expressed above is equal to the so-called fiscal externality associated with these taxes.[7] We will return to this important observation when discussing environments in which governments may use other taxes to achieve their objectives, and show that the same Pigouvian logic continues to apply.

Atkin and Donaldson (2022) provide estimates of these misallocation terms for various countries around the world, which one could use again to conduct optimal trade policy. If import-competing sectors tend to have lower markups, then the optimal trade policy may be import subsidies that relocate workers towards sectors with higher markups. The same logic also applies when the "labor wedge" between the social and private marginal costs of employment arises from nominal wage rigidities, as emphasized by Johnson (1965).

Trade Policy as a Tool for Domestic Redistribution

Traditionally, Pigou's ideas have been associated with nonpecuniary considerations. An inefficient level of economic activities, perhaps pollution or employment, is what creates a rationale for taxation. Interpreted more broadly, the same ideas, however, can be equally useful to shed light on pecuniary considerations. We start by discussing domestic redistribution here.

Suppose now that the allocation of resources in the economy is efficient and that international prices are fixed, so that the only relevant consideration for evaluating the effects of trade policy is how it affects domestic prices $(p_1, \ldots, p_N)$. This long vector includes the prices of all tradable goods that may directly be affected by trade policy, as well as the prices of nontradable goods, wages, and any other primary factor of production that may also be indirectly affected by trade policy. In this case,

[7] Harberger (1964) offers an early discussion of the equivalence between taxes and various distortions or wedges. More recently, this equivalence plays a central role in the misallocation literature, from Hsieh and Klenow (2009) to Baqaee and Farhi (2020).

going through the same hypothetical scenario where the government lowers the imports of good g, the necessary condition for the observed tariff on good g to be optimal is

$$\text{(Domestic Redistribution)} \qquad t_g = \sum_n \textit{Social Harm}\left(p_n\right) \times \frac{\Delta p_n}{\Delta m_g},$$

with the social harm of raising p_n now a combination of the social marginal cost for the buyers of good n and the social marginal benefit for their sellers,

$$\textit{Social Harm}\left(p_n\right) = \sum_h \textit{Social Return of Fiscal Transfer to } h \times m_n\left(h\right),$$

where $m_n(h)$ denotes net trade of a good or factor n by a household h. Net trade $m_n(h)$ is positive if h is a net buyer of n (for example, of electric vehicles or t-shirts) and negative if h is a net seller (for example, of blue- or white-collar labor services).

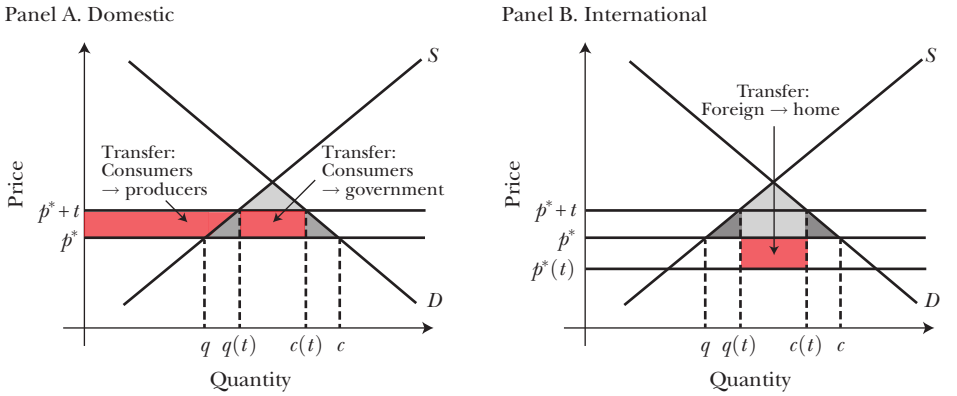
If the marginal value of a dollar in the hands of the government is the same as the marginal value of a dollar in the hands of household h, then the social return of a fiscal transfer to h is zero. And if this is true for all households, then the redistributive tariff should be set equal to zero. This assumption was implicit in the classical case for free trade described in Figure 1. But in general, one may value using trade policy to affect domestic prices and, in turn, the real income of different households. The domestic redistribution formula plays a central role in the political-economy-of-trade literature, with Grossman and Helpman (1994) offering the most influential micro-foundations for differences in the social return of fiscal transfers to different households in a model of lobbying.

Figure 3, panel A, offers a graphical representation where we have highlighted (in red) the transfers from consumers to producers and from consumers to the governments.[8] For a given social welfare criterion, one can view the fact that the social marginal value of income is not equalized across households as another source of distortion. In the same way that policy intervention could improve welfare by reallocating resources to the "right" activities—that is, those deemed to have higher marginal values—policy intervention may now improve welfare by reallocating income to the "right" households, even if production were to remain unchanged.

Using tariffs to finance an exogenous stream of government revenues can be viewed as a special case of domestic redistribution, from the private sector to the government. In this case, the magnitude of government expenditure determines the (negative) social return of transfers to a representative domestic household. The domestic redistribution formula for an optimal tariff then reduces to Ramsey's tax formula, which calls for minimizing the real income cost to the domestic household, subject to satisfying the budgetary needs of the government.

[8] To make the redistributive effects stark, we treat consumers and producers as distinct groups of households in this example. Consumers own the outside good, which they sell to purchase the good whose demand and supply are displayed in Figure 3, panel A. Producers own the output of this good, which they sell to purchase the outside good.

Figure 3
Trade Policy as a Tool for Redistribution



Source: Authors' calculations.
Note: Panel A describes the domestic transfers associated with an import tariff. Panel B describes the international transfers associated with an import tariff.

As in the case of trade policy as a tool for efficiency, whether or not there are general-equilibrium effects does not affect the previous Pigouvian logic. It only affects how a government would go about restricting imports to achieve its redistribution objective. For instance, if there are no general-equilibrium effects, a government that wants to redistribute income towards textile workers may only want to restrict textile imports. If there are general-equilibrium effects, a government should also take into account how input-output linkages and factor mobility across sectors would affect the return to these workers. In this case, import restrictions on one good may have implications for the prices of other goods and factors. Quantitative trade models and computers may again be required, as in Adão, Becko, Costinot, and Donaldson (2025), to solve for these complicated effects and derive estimates of $\Delta p_n/\Delta m_g$ for the domestic redistribution formula. But the Pigouvian logic remains straightforward: if $\Delta p_n/\Delta m_g > 0$ and the social harm of raising p_n is positive, because it lowers the real income of households with higher social marginal value of income, then the optimal tariff on good g should be set equal to the social harm that it generates. When poor households have higher social marginal value of income than rich households, trade policy can raise social welfare by engineering, through changes in domestic prices, a progressive redistribution of income.

Trade Policy as a Tool for International Redistribution

We can apply the Pigouvian logic in the exact same manner when the motive for trade policy intervention is international redistribution. Suppose again that the domestic allocation of resources is efficient, but now assume that by restricting imports, a country may affect the international prices $(p_1^*, \ldots, p_N^*)$ paid and received by foreigners. Again, this long vector may include the prices of tradable goods,

but also the prices of nontradable goods and wages.[9] Finally, in order to isolate the international redistribution motive, suppose that the social return of a fiscal transfer to any domestic household is zero.

Let us return to a hypothetical policy reform that reduces imports of good g and, in turn, triggers an adjustment in the vector of international prices. Except for the fact that international rather than domestic prices are now the variable of interest, there is no difference compared to the case that we have just studied. For the observed tariff on good g to be optimal, the following condition must hold

$$\text{(International Redistribution)} \qquad t_g = \sum_n \textit{Social Harm}\left(p_n^*\right) \times \frac{\Delta p_n^*}{\Delta m_g}.$$

with the social harm of raising a given price p_n^* a function of whether foreigners are net buyers or sellers of good n,

$$\textit{Social Harm}\left(p_n^*\right) = \textit{Social Return of Fiscal Transfer to Foreigners} \times m_n^*,$$

where $m_n^* = -m_n$ are net foreign imports of good n, in line with our previous notation for domestic households.

The social return of a fiscal transfer to foreigners depends on the social marginal value of foreigners' income. The classical optimal tariff argument, discussed in the early work of Mill (1844), Torrens (1844), and later Edgeworth (1894), assumes that the social marginal value of foreigners' income is exactly zero. Under this assumption, the social return of a fiscal transfer to them is equal to -1. In this case, the international redistribution tariff formula simplifies into

$$\text{(Terms-of-Trade Manipulation)} \qquad t_g = \sum_n m_n \times \frac{\Delta p_n^*}{\Delta m_g}.$$

This formulation captures the idea that a country with the ability to manipulate its terms of trade may exercise its monopsony power to reduce the price of its imports and its monopoly power to raise the price of its exports, thereby transferring income from foreigners to itself. The terms-of-trade manipulation formula therefore mimics the optimal pricing rule of a multi-good firm that acts both as a monopsonist and as a monopolist. In the special case with only two goods, with good g the one subject to an import tariff and the other good the numeraire, the previous formula implies that the optimal import tariff, expressed in ad-valorem terms, should be equal to the inverse of the elasticity of the foreign export supply curve (that is, $t_g/p_g^* = \Delta \ln p_g^*/\Delta \ln m_g$), as in Johnson (1950).

[9] In theory, this long vector could also include the prices of goods in different periods or states of the world. In Costinot, Lorenzoni, and Werning (2014), we use this broader interpretation to tackle the question of optimal capital controls as a special case of optimal trade policy. Dávila, Rodríguez-Clare, Schaab, and Tan (2025) use a similar approach to study optimal tariffs in dynamic heterogeneous agent economies.

Figure 3, panel B, offers a graphical representation. The upside of distorting imports of good g is to lower its international price and create a transfer from foreigners (in red). If a country cannot affect international prices, then the terms-of-trade manipulation formula implies free trade. But if it can, and domestic consumers and firms do not internalize the impact of their behavior on international prices, Pigou's logic calls for an optimal tax on each good g that is equal to the marginal effect of its imports m_g on the country's terms of trade, $\sum_n m_n \times \Delta p_n^*$.[10]

More generally, the social marginal value of foreigners' income could be either positive (perhaps because of international cooperation along other dimensions, as in Adão, Becko, Costinot, and Donaldson 2026) or negative (perhaps because of the desire to impose economic sanctions, as in Becko 2024). A lower value implies a lower social fiscal return of transfer to foreigners and, in turn, a uniform shift in the optimal trade policy t_g, but it does not affect how trade taxes should vary across goods, as emphasized by Becko (2024). General-equilibrium considerations, as complex as they might be, are again encapsulated in the marginal effect of import restrictions on the distorted economic variables, now in the form of the vector of international prices. The emergence of global value chains, a much discussed phenomenon, matters if it affects these marginal effects.[11]

We have abstracted so far from the issue of foreign retaliation. Would retaliation invalidate the argument? As a theoretical matter, the answer is "no." These considerations would affect the value of world price changes $\Delta p_n^*/\Delta m_g$ entering the international redistribution formula, but not the structure of the formula itself. Suppose that changes in trade policy at home cause changes in trade policy in the rest of the world. In this case, foreign retaliation, whether conducted by one or many trading partners, becomes one more determinant of the response of international prices to imports, alongside the other economic considerations that shape those price adjustments. Likewise, a country could anticipate that under World Trade Organization rules, any change in tariffs would be reciprocated by a change in foreign tariffs that keeps its terms of trade fixed, as in Bagwell and Staiger (1999). As a practical matter, of course, retaliation is a potentially important argument

[10] We have highlighted the parallel between the domestic and international redistribution formulas. A similar connection holds between the misallocation and international redistribution formulas. When international prices are endogenous, domestic consumers and firms do not use the trade technology efficiently because they fail to recognize that the social marginal cost of their imports m_g includes the price effects on infra-marginal units of all other goods. The optimal Pigouvian tax asks them to pay for $\textit{Social Harm}(m_g) = \sum_n m_n \times \frac{\Delta p_n^*}{\Delta m_g}$, consistent with the misallocation formula.

[11] An interesting but distinct issue related to global value chains is whether they lead to cross-country ownership of assets. Implicit in our discussion is the assumption that foreigners face international prices p^* and not domestic prices p. If foreigners own domestic assets because of foreign direct investment, then changes in p would also affect the magnitude of income transfers to the rest of the world. Blanchard (2009) offers an example. In such situations, a country may have incentives to lower domestic prices uniformly as a way to expropriate foreigners, leading to departures of Lerner Symmetry, as we discuss in Costinot and Werning (2019).

against attempts at terms-of-trade manipulation, as the foreign policy change could even turn what would have been a positive transfer from abroad into a negative one.

How large are the tariffs called for by the terms-of-trade manipulation formula? Quantitative trade models can provide values under assumptions about the structure of preferences, technology, and market structure. In Ossa (2014), abstracting from foreign retaliation, the median values are around 60 percent. In Broda, Limão and Weinstein (2008), using a simpler partial equilibrium analysis, they are around 160 percent! None of these estimates, however, use actual changes in trade policy to estimate its incidence on international prices directly. Instead, incidence is inferred indirectly, through the lens of a structural model, via the response of import prices and quantities to other shocks. The question of whether countries can manipulate their terms of trade—and in turn let foreign exporters pay for tariffs via a decrease in the prices they receive—is far from settled empirically. Recent evidence collected during the US-China trade war points to perhaps surprisingly small responses of international prices to US tariffs. Both Fajgelbaum, Goldberg, Kennedy, and Khandelwal (2020) and Amiti, Redding and Weinstein (2019) cannot reject a zero effect of US tariffs on the international prices of US imports that were targeted by tariffs, relative to US imports that were not. Put differently, there was almost complete pass-through of US tariffs into domestic prices. This is interesting, but it should be clear that the results from these difference-in-difference regressions do not imply that the optimal tariff is zero. The terms-of-trade manipulation formula includes all goods, including those that the United States exports. The United States could improve its terms of trade if it could raise the price of its exports relative to its imports, but recent empirical evidence simply does not shed light on the size of this effect.

Act IV: The Targeting Principle

Compared to the classical case for free trade, which assumes that laissez-faire is optimal, the Targeting Principle offers a more subtle and robust defense. The economy may be distorted and government intervention may be warranted. Yet one cannot conclude that a tariff is optimal without making it compete with other available instruments. The Targeting Principle suggests that trade policy is acupuncture with a fork and calls for using needles instead.

A Better Way to Improve Efficiency

To put a little more meat on the distinction between forks and needles, let us return first to the case of a domestic externality e, such as air pollution or local knowledge spillovers. Suppose that the government, in addition to trade policy, has access to another policy instrument t^e, which it can use to tax or subsidize e. This will be the needle in this example.

A simple way to understand how the Targeting Principle operates is to note that with access to another policy instrument, it is optimal for the government to set

$t^e = \textit{Social Harm}(e)$, which is the textbook Pigouvian tax that restores the first-best allocation. Now returning to the externality formula, and taking into account the fiscal externality associated with the corrective tax $t^e \Delta e$, this implies

$$t_g = \left[\textit{Social Harm}(e) - t^e\right] \times \frac{\Delta e}{\Delta m_g} = 0.$$

When the other policy is in place, there is no longer a wedge between private and social marginal costs and the rationale for using non-zero trade taxes evaporates.

Intuitively, when trade policy is used to correct a distortion, it is at the cost of potentially creating another. In Figure 2, for instance, the problem is the level of domestic output ($e = q$). A small tariff helps to bring q closer to the social optimum, but at the cost of pushing domestic consumption c away from its optimal level. In this example, if a production subsidy is available, it ought to be used to target the underlying distortion. This is why trade policy intervention is acupuncture with a fork.

The same logic applies if there are multiple domestic externalities, if some of these externalities are dynamic, as in the case of the infant industry argument, or if there is domestic misallocation due to imperfect competition, nominal rigidities, or financial constraints. If the government has access to additional policy instruments to target domestic distortions directly, then it should take advantage of these and let trade be free.

A Better Way to Redistribute

The Targeting Principle can also be applied to adjudicate the role that trade policy should play, or not, as an instrument for domestic redistribution, either across different households or as a way to collect revenues for government spending.

As a first benchmark, suppose that the government has access to unrestricted lump-sum transfers, an admittedly extreme assumption. In this environment, lump-sum transfers should be engineered up to the point at which the social marginal utility of income of every household is equal to the social marginal value of public funds and the marginal return of a fiscal transfer is zero. Because lump-sum transfers are nondistortionary, our domestic redistribution formula continues to apply. But once the social return of fiscal transfers is zero, the formula implies zero trade taxes. The alternative policy tool means that seeking to use trade for redistribution is superfluous, as lump-sum transfers alone restore the first-best allocation.

Of course, unrestricted lump-sum transfers are not available in practice, and it is hard to imagine that they ever will be. As a first step towards greater realism, Diamond and Mirrlees (1971) and Dixit and Norman (1986) turn to an environment in which the government instead has access to a complete set of linear commodity taxes t^c, which it could use to pursue its redistribution objective. The set of commodity taxes include both taxes on goods and factors, which can create a wedge between the prices faced by domestic households and domestic firms. Would the government also want to use trade policy in this case? A corollary of Diamond and Mirrlees' (1971) production efficiency result is that trade, viewed as another technology to deliver domestic consumption, should not be taxed.

To understand how this celebrated result relates to the Targeting Principle and the Pigouvian logic developed earlier, consider again a hypothetical policy reform that reduces imports of good g by some small amount, but now accompanied by a change in commodity taxes that keeps fixed the prices p faced by domestic households. A complete set of commodity taxes t^c makes such a combined policy reform feasible. Because domestic prices are fixed, there are no income transfers across households and the optimal tariff implied by the domestic redistribution formula would again be zero. The remaining issue is whether commodity taxation creates additional distortions. The answer is no, as households have no reason to alter their behavior. It follows again that the optimal policy must be free trade.[12]

Even when optimal commodity taxes are in place, the economy considered by Diamond and Mirrlees (1971) and Dixit and Norman (1986) does not replicate the first-best allocation. When the government's goal is to redistribute, rather than improve efficiency, it must introduce distortions that would have been absent under laissez-faire. The more robust insight here is that if needles exist to achieve a given policy goal—in this case, in the form of commodity taxes that can directly target the prices faced by domestic households—then they should be preferred to coarser trade policy instruments.

Act V: Two Limits to the Targeting Principle?

As a defense of free trade, the Targeting Principle has two limits. First, it only renders trade policy intervention superfluous when the relevant targets are domestic rather than international. Second, even when the policy target is domestic, the needles that it calls for may be much finer than those available to governments in practice, in which case tariff policy may reappear on the policy menu as a flawed but practical real-world alternative.

When Targets Are International

The classical optimal tariff argument, which emphasizes a country's ability to manipulate its terms of trade, is often presented as the key theoretical exception to the optimality of free trade. Irwin (1996) describes it as "the hardest to refute on theoretical grounds and [. . .] the most important exception to free trade ever

[12] For the curious and mathematically inclined readers, the complete argument is as follows. Recall that if there are multiple rationales for trade policy intervention, then one should take the sum of our Pigouvian formulas. In the presence of distortionary commodity taxation, this implies

$$t_g = -\sum_n t_n^c \times \frac{\Delta c_n}{\Delta m_g} + \sum_n \textit{Social Harm}(p_n) \times \frac{\Delta p_n}{\Delta m_g},$$

where the fiscal externality $-\sum_n t_n^c \times (\Delta c_n / \Delta m_g)$ reflects the efficiency motive in the misallocation formula. The policy reform is constructed so that $\Delta p_n = \Delta c_n = 0$ for all n. This generalized formula therefore implies $t_g = 0$.

conceived." One reason is that trade policy intervention in this case is fully consistent with the Targeting Principle.

From an empirical standpoint, one may question the magnitude of terms-of-trade movements that an individual country is able to engineer, especially in the presence of foreign retaliation. From a moral standpoint, one may also object to a government assigning zero marginal value to income in the rest of the world and, in turn, benefiting from making foreigners poorer. Irwin (1991) reminds us of the memorable words of Edgeworth (1908), who helped lay the theoretical foundations of the terms-of-trade argument. Referring to the subsequent work of Bickerdike (1906), he notes: "The direct use of the theory is likely to be small. But it is to be feared that its abuse will be considerable . . . Let us admire the skill of the analyst, but label the subject of his investigation POISON."

However, if a government's objective is international redistribution, it should be clear that trade policy is a targeted way to achieve it. As previously discussed, the classical tariff argument, formalized in the terms-of-trade manipulation formula, corresponds to the special case where the targets of interest are international prices. There are no domestic policy instruments that can give a country direct control over those prices, above and beyond what trade policy can achieve.

This observation applies more broadly to any international target, not just international prices. We have already mentioned the case of global carbon emissions and geopolitical actions. What these two examples have in common is that the externality derives, at least partly, from actions taken in the rest of the world, which domestic policy instruments do not directly control. Hence, a country that decides to fight climate change unilaterally should tax its own carbon emissions. But because a country cannot directly tax foreign carbon emissions, trade policy is, from its perspective, the optimal way to tackle emissions in the rest of the world. Compared to the earlier externality formula, the only difference is that the relevant changes in carbon emissions are foreign rather than global ones. If foreign emissions arise solely from domestic consumption, the optimal tariff on good g reduces to the social cost of carbon multiplied by the carbon content of good g's imports, consistent with a carbon border adjustment mechanism. If foreign emissions partly derive from foreign consumption, then optimal trade policy ought to seek to affect these emissions indirectly as well.

Compared to the classical tariff argument, terms-of-trade manipulation is no longer an end in itself, but it remains critical. If countries cannot affect international prices, then they are unlikely to affect carbon emissions in the rest of the world either, as it is changes in international prices that incentivize beef producers in the Amazon or palm oil producers in Indonesia to alter their behavior.[13] The same logic applies to tariffs used as sanctions for geopolitical purposes.

[13]A small caveat is that our observation presumes a unique mapping from international prices to foreign exports and imports. In theory, foreign firms may be indifferent between selling domestically or exporting, in which case trade policy in one country may affect quantities in the rest of the world,

When Domestic Instruments Are Coarse

Let us return once more to the case where domestic redistribution is the only motive for trade policy intervention. On the one hand, it is unrealistic to assume that no other policy instruments are available to achieve this goal. As Rodrik (1995) puts it, "Saying that trade policy exists because it serves to transfer income to favored groups is a bit like saying Sir Edmund Hillary climbed Mt. Everest because he wanted to get some mountain air. There was surely an easier way of accomplishing that objective." On the other hand, it is equally unrealistic to assume the availability of lump-sum transfers to achieve redistribution at no efficiency cost.

The production efficiency result of Diamond and Mirrlees (1971) discussed earlier relies on assumptions that are less extreme than the availability of unrestricted lump-sum transfers, but still demanding. For their result to hold, the government needs access to a complete set of commodity taxes—not just different taxes on different goods, such as electric vehicles and tee-shirts, but also on different factors, such as tech and textile workers. In principle, workers with distinct skill sets may be taxed differently, even if they earn the same income.

In Costinot and Werning (2023), we explore a more realistic environment in which the government can use domestic taxes for redistribution, but these are limited to nonlinear income taxes, as in Naito (1999). The only restriction that we impose on social preferences is that the government cares about the distribution of real income, not the identity of who is rich or poor.[14] In this setting, we show that trade policy can be used as a form of predistribution that complements redistribution through income taxation.

To see how departures from free trade may arise, consider again a reduction in imports of good g, now accompanied by an adjustment in the income tax schedule designed to hold fixed the distribution of real income. By construction, there is no direct benefit to this policy reform in terms of redistribution. But unlike in the case of commodity taxation discussed above, there is now potentially a non-zero fiscal externality arising from changes in labor supply. By the same Pigouvian logic as in the misallocation formula, the optimal trade tax on good g must be equal to

$$t_g = -\sum_n \tau_n \times \left(w_n \ell_n\right) \times \frac{\Delta \ln \ell_n}{\Delta m_g},$$

where τ_n is the marginal income tax rate faced by households at the n-th quantile of the wage distribution, $w_n \ell_n$ is their earnings, and $\Delta \ln \ell_n$ is the resulting change

even without changes in international prices. Such a situation arises in Kortum and Weisbach's (2021) Ricardian model. We thank Sam Kortum for pointing this out.

[14] This anonymity criterion is attractive from a normative standpoint, our primary focus in this paper. In practice, however, governments may place greater weight on individuals associated with certain sectors or states, perhaps because of differences in their ability to lobby or to be pivotal in elections. If so, their choice of trade taxes would also reflect these political-economy considerations, in line with the domestic redistribution formula. Adão, Becko, Costinot, and Donaldson (2025) conclude that such considerations account for about one-third of US tariff variation.

in labor supply. In addition, in order to keep real income fixed, changes in labor supply must offset changes in relative wages, implying

$$\frac{\Delta \ln \ell_n}{\Delta m_g} \propto -\frac{\Delta \ln \omega_n}{\Delta m_g},$$

where $\omega_n = w_{n+1} - w_n/w_n$ measures wage growth across quantiles, that is, inequality, and the coefficient of proportionality is pinned down by labor supply elasticities ($\propto$ means "proportional to").

When relative wages are fixed, so that $\Delta \ln \omega_n / \Delta m_g = 0$, this new formula implies free trade. But when imports are associated with a steepening of the wage schedule, so that $\Delta \ln \omega_n / \Delta m_g > 0$, the same formula calls for trade policy to be used as a form of predistribution. Unlike in the domestic redistribution formula, knowledge of the social marginal value of transfers to different households is not required. This opens up the possibility of a purely empirical evaluation of whether observed trade policy is optimal.

As an illustration of how to combine theory and data to shed light on optimal trade policy, Costinot and Werning (2023) use existing empirical work on the China shock as inputs into the above formula. A large literature has documented the effects of the rise of China on US labor markets, most notably Autor, Dorn, and Hanson (2013), Autor, Dorn, Hanson, and Song (2014), and Acemoglu et al. (2016). This important body of work, however, leaves open the question of what these effects imply for US policy. Using estimates of the effect of Chinese import competition on US wages at different quantiles of the wage distribution from Chetverikov, Larsen, and Palmer (2016), along with measures of marginal US income tax rates and estimates of labor supply elasticities, this formula implies an import tariff on Chinese goods of 0.07 percent.

In a nutshell, the available evidence suggests that large changes in imports induce modest changes in relative wages, in the sense that $\Delta \ln \omega_n / \Delta m_g$ is small. As a result, the fiscal cost of distorting imports in order to achieve predistribution is high and the optimal tariff is low. Although free trade need not be optimal in this context, it remains an excellent approximation.

Concluding Remarks

There is more to optimal trade policy than free trade. Mercantilist fallacies aside, we have shown that the same Pigouvian perspective can be used as a unifying framework to understand when trade policy intervention is warranted and when it is not. From the classical optimal tariff argument of Mill (1844) and Torrens (1844) to contemporary debates about global carbon emissions and geopolitics, all the formulas presented in this article derive from the same basic principle: if domestic firms and consumers do not internalize the marginal impact of their imports on social welfare, they should be asked to pay for it. Despite the complexity of general-equilibrium interactions, optimal trade policy can be expressed as

a function of a few sufficient statistics that do not depend on the details of such interactions.

We recognize that many of these statistics may be hard to measure in practice and unlikely to be pinned down by empirical work alone. For instance, if the European Union were to reduce its imports of Brazilian beef, by how much would carbon emissions abroad decline? Current answers to such questions depend not only on data, but also on the specific structure of the quantitative trade models used to harness them.

Yet such limitations do not lessen the importance of clarifying when and why a government may wish to tax or subsidize trade. The Pigouvian formulas presented in this article highlight what matters and what ideally needs to be known. To go from data to policy recommendation, one needs a theoretical compass. We hope this article helps orient ongoing and future debates in the right direction.

■ *The authors thank John Becko, Dave Donaldson, Elhanan Helpman, Doug Irwin, Sam Kortum, Ben Olken, and Andres Rodriguez-Clare for helpful comments. Arnaud Costinot also gratefully acknowledges the financial support of NSF grant number 2242367.*

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.

Adão, Rodrigo, John Sturm Becko, Arnaud Costinot, and Dave Donaldson. 2025. "Why Is Trade Not Free? A Revealed Preference Approach." NBER Working Paper 31798.

Adão, Rodrigo, John Sturm Becko, Arnaud Costinot, and Dave Donaldson. 2026. "World Trading System for Whom? Evidence from Global Tariffs." NBER Working Paper 34658.

Amiti, Mary, Stephen J. Redding, and David E. Weinstein. 2019. "The Impact of the 2018 Tariffs on Prices and Welfare." *Journal of Economic Perspectives* 33 (4): 187–210.

Atkin, David, and Dave Donaldson. 2022. "The Role of Trade in Economic Development." In *Handbook of International Economics*, Vol. 5, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 1–59. Elsevier.

Auclert, Adrien, Matthew Rognlie, and Ludwig Straub. 2025. "The Macroeconomics of Tariff Shocks." NBER Working Paper 33726.

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

Autor, David H., David Dorn, Gordon H. Hanson, and Jae Song. 2014. "Trade Adjustment: Worker-Level Evidence." *Quarterly Journal of Economics* 129 (4): 1799–860.

Bagwell, Kyle, and Robert W. Staiger. 1999. "An Economic Theory of GATT." *American Economic Review* 89 (1): 215–48.

Baqaee, David Rezza, and Emmanuel Farhi. 2020. "Productivity and Misallocation in General Equilibrium." *Quarterly Journal of Economics* 135 (1): 105–63.

Bardhan, Pranab K. 1970. *Economic Growth, Development, and Foreign Trade*. Wiley.

Barwick, Panle Jia, Hyuk-Soo Kwon, Shanjun Li, and Nahim B. Zahur. 2025. "Drive Down the Cost: Learning by Doing and Government Policies in the Global EV Battery Industry." NBER Working Paper 33378.

Becko, John Sturm. 2024. "A Theory of Economic Sanctions as Terms-of-Trade Manipulation." *Journal of International Economics* 150: 103898.

Becko, John S., Gene M. Grossman, and Elhanan Helpman. 2025. "Optimal Tariffs with Geopolitical Alignment." NBER Working Paper 34108.

Becko, John Sturm, and Daniel G. O'Connor. 2025. "Strategic (Dis)Integration." Unpublished.

Beveridge, William. 1931. *Tariffs: The Case Examined*. Longmans.

Bhagwati, Jagdish, Ronald Jones, Robert Mundell, and Jaroslav Vanek, eds. 1971. *Trade, Balance of Payments, and Growth: Papers in International Economics in Honor of Charles P. Kindleberger*. North-Holland.

Bickerdike, Charles F. 1906. "The Theory of Incipient Taxes." *Economic Journal* 16 (64): 529–35.

Blanchard, Emily J. 2009. "Trade Taxes and International Investment." *Canadian Journal of Economics/ Revue canadienne d'économique* 42 (3): 882–99.

Broda, Christian, Nuno Limão, and David Weinstein. 2008. "Optimal Tariffs and Market Power: The Evidence." *American Economic Review* 98 (5): 2032–65.

Caliendo, Lorenzo, Samuel S. Kortum, and Fernando Parro. 2025. "Tariffs and Trade Deficits." NBER Working Paper 34003.

Chetverikov, Denis, Bradley Larsen, and Christopher Palmer. 2016. "IV Quantile Regression for Group-Level Treatments, with an Application to the Distributional Effects of Trade." *Econometrica* 84 (2): 809–33.

Clayton, Christopher, Matteo Maggiori, and Jesse Schreger. 2026. "A Framework for Geoeconomics." *Econometrica* 94 (1): 105–36.

Corden, W. Max. 1974. *Trade Policy and Economic Welfare*. Clarendon Press.

Costinot, Arnaud, Guido Lorenzoni, and Ivan Werning. 2014. "A Theory of Capital Controls as Dynamic Terms-of-Trade Manipulation." *Journal of Political Economy* 122 (1): 77–128.

Costinot, Arnaud, and Andres Rodríguez-Clare. 2018. "The US Gains from Trade: Valuation Using the Demand for Foreign Factor Services." *Journal of Economic Perspectives* 32 (2): 3–24.

Costinot, Arnaud, and Andres Rodríguez-Clare. 2025. "Seven Questions about Tariffs That Everyone Should Know the Answer To." In *The Economic Consequences of The Second Trump Administration: A Preliminary Assessment*, edited by Gary Gensler, Simon Johnson, Ugo Panizza, and Beatrice Weder di Mauro. CEPR Press.

Costinot, Arnaud, and Iván Werning. 2019. "Lerner Symmetry: A Modern Treatment." *American Economic Review: Insights* 1 (1): 13–26.

Costinot, Arnaud, and Iván Werning. 2023. "Robots, Trade, and Luddism: A Sufficient Statistics to Optimal Technology Regulation." *Review of Economic Studies* 90 (5): 2261–91.

Costinot, Arnaud ⓒ Iván Werning. 2025. "How Tariffs Affect Trade Deficits." NBER Working Paper 33709.

Dávila, Eduardo, Andrés Rodríguez-Clare, Andreas Schaab, and Stacy Tan. 2025. "A Dynamic Theory of Optimal Tariffs." NBER Working Paper 33898.

Diamond, Peter A., and James A. Mirrlees. 1971. "Optimal Taxation and Public Production I: Production Efficiency." *American Economic Review* 61 (1): 8–27.

Dixit, Avinash. 1985. "Tax Policy in Open Economies." In *Handbook of Public Economics*, Vol. 1, edited by Alan J. Auerbach and Martin Feldstein, 313–74. Elsevier.

Dixit, Avinash, and Victor Norman. 1986. "Gains from Trade without Lump-Sum Compensation." *Journal of International Economics* 21 (1–2): 111–22.

Eaton, Jonathan, Samuel Kortum, and Brent Neiman. 2016. "Obstfeld and Rogoff's International Macro Puzzles: A Quantitative Assessment." *Journal of Economic Dynamics and Control* 72: 5–23.

Edgeworth, Francis Y. 1894. "Theory of International Values." *Economic Journal* 4 (13): 35–50.

Edgeworth, Francis Y. 1908. "Appreciations of Mathematical Theories." *Economic Journal* 18 (72): 541–56.

Fajgelbaum, Pablo D., Pinelopi K. Goldberg, Patrick J. Kennedy, and Amit K. Khandelwal. 2020. "The Return to Protectionism." *Quarterly Journal of Economics* 135 (1): 1–55.

Farrokhi, Farid, and Ahmad Lashkaripour. 2025. "Can Trade Policy Mitigate Climate Change?" *Econometrica* 93 (5): 1561–99.

Garci-Lembergman, Ezequiel, Natalia Ramondo, Andres Rodríguez-Clare, and Joseph S. Shapiro. 2025. "Carbon Taxes in the Global Economy: A Quantitative Analysis." Unpublished.

Goldberg, Pinelopi K., Réka Juhász, Nathan J. Lane, Giulia Lo Forte, and Jeff Thurk. 2024. "Industrial Policy in the Global Semiconductor Sector." NBER Working Paper 32651.

Grossman, Gene M., and Elhanan Helpman. 1994. "Protection for Sale." *American Economic Review* 84 (4): 833–50.

Harberger, Arnold C. 1964. "The Measurement of Waste." *American Economic Review* 54 (3): 58–76.

Helpman, Elhanan, and Paul R. Krugman. 1989. *Trade Policy and Market Structure*. MIT Press.

Hsiao, Allan. 2026. "Coordination and Commitment in International Climate Action: Evidence from Palm Oil." *Econometrica* 94 (1): 1–33.

Hsieh, Chang-Tai, and Peter J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *Quarterly Journal of Economics* 124 (4): 1403–48.

Irwin, Douglas A. 1991. "Retrospectives: Challenges to Free Trade." *Journal of Economic Perspectives* 5 (2): 201–08.

Irwin, Douglas A. 1996. *Against the Tide: An Intellectual History of Free Trade*. Princeton University Press.

Irwin, Douglas A. 2012. *Trade Policy Disaster: Lessons from the 1930s*. MIT Press.

Irwin, Douglas A., and Peter J. Klenow. 1994. "Learning-by-Doing Spillovers in the Semiconductor Industry." *Journal of Political Economy* 102 (6): 1200–227.

Johnson, Harry G. 1950. "Optimum Welfare and Maximum Revenue Tariffs." *Review of Economic Studies* 19 (1): 28–35.

Johnson, Harry G. 1965. "Optimal Trade Intervention in the Presence of Domestic Distortions." In *Trade, Growth and the Balance of Payments*, edited by Robert Baldwin. North-Holland.

Kortum, Samuel, and David A. Weisbach. 2021. "Optimal Unilateral Carbon Policy." Cowles Foundation Discussion Paper 2311.

Krugman, Paul. 1996. *Pop Internationalism*. MIT Press.

Lerner, Abba P. 1936. "The Symmetry between Import and Export Taxes." *Economica* 3 (11): 306–13.

Lipsey, R. G., and Kevin Lancaster. 1956. "The General Theory of Second Best." *Review of Economic Studies* 24 (1): 11–32.

Meade, James E. 1955. *The Theory of International Economic Policy*. Vol. 2. Oxford University Press.

Mill, John Stuart. 1844. *Essays on Some Unsettled Questions of Political Economy*. Parker.

Mun, Thomas. 1664. *England's Treasure by Forraign Trade*. T. Clark.

Naito, Hisahiro. 1999. "Re-examination of Uniform Commodity Taxes under Non-linear Income Tax System and Its Implications for Production Efficiency." *Journal of Public Economics* 71 (2): 165–88.

Obstfeld, Maurice, and Kenneth Rogoff. 2001. "The Six Major Puzzles in International Macroeconomics: Is There a Common Cause?" In *NBER Macroeconomics Annual 2000*, Vol. 15, edited by Ben S. Bernanke and Kenneth Rogoff, 339–412. MIT Press.

Ossa, Ralph. 2014. "Trade Wars and Trade Talks with Data." *American Economic Review* 104 (12): 4104–46.

Reyes-Heroles, Ricardo. 2016. "The Role of Trade Costs in the Surge of Trade Imbalances." Unpublished.

Ricardo, David. 1817. *The Principles of Political Economy and Taxation*. John Murray.

Rodrik, Dani. 1995. "Political Economy of Trade Policy." In *Handbook of International Economics*, Vol. 3, edited by Gene M. Grossman and Kenneth Rogoff, 1457–94. Elsevier.

Tinbergen, Jan. 1952. *On the Theory of Economic Policy*. North-Holland Publishing Company.

Torrens, Robert. 1844. *The Budget: On Commercial and Colonial Policy*. Smith, Elder.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 123–144

The Incidence of Tariffs: Rates and Reality

Gita Gopinath and Brent Neiman

Statutory tariff rates on US imports have risen dramatically to levels that have not been seen in more than 100 years. At the end of 2025, the trade-weighted average statutory rate sat at 26 percent. Imports from 171 exporting nations, on goods accounting for more than 70 percent of total US imports, faced higher tariffs than they did at the end of 2024.

So far, retaliation has been limited, although US announcements of large and broad tariffs on imports from China were met with Chinese announcements of comparably large and broad tariffs on imports from the United States. The back-and-forth led at one point to obviously prohibitive tariff rates that exceeded 100 percent and covered almost all trade between the countries. The United States and China subsequently reduced their bilateral tariff rates, but they remain historically high. The peak of the storm may have passed, but as Dorothy observed in *The Wizard of Oz*, "We're not in Kansas anymore."

What consequence might we expect from this surge of tariffs? First, we might expect that the United States will pay higher prices for affected imports. The impact on US import prices will depend not only on the size of the tariffs, but also on the degree of tariff pass-through, which could be low if foreign producers reduce export prices or high if they leave export prices unchanged. Second, there will be a shift in US import spending away from goods with the largest price increases and toward other foreign or domestic suppliers. Third, most models predict that the

■ *Gita Gopinath is Professor of Economics, Harvard University, Cambridge, Massachusetts. Brent Neiman is Professor of Economics, Booth School of Business, University of Chicago, Chicago, Illinois. Their email addresses are gopinath@fas.harvard.edu and brent.neiman@chicagobooth.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20251486.

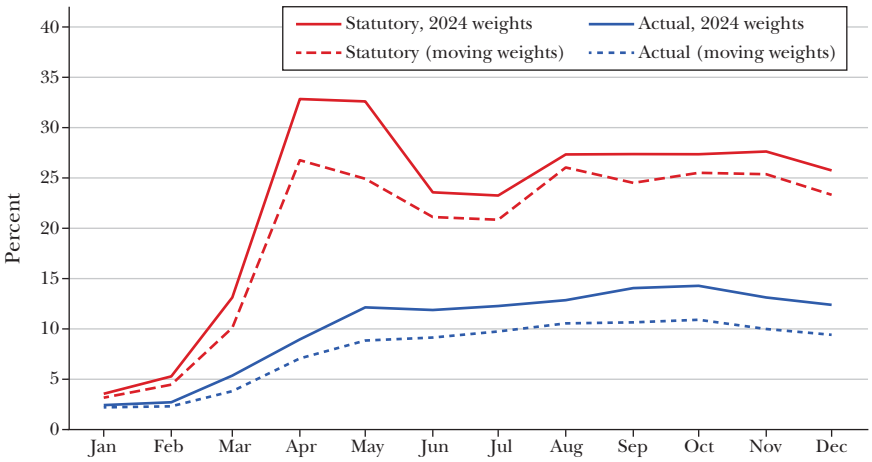
tariffs should lead to an appreciation of the US dollar, which would mute the rise in import prices but impair demand for US exports. In this paper, we examine the effects of US import tariffs enacted during 2018–2019 and 2025, with a focus on these three predictions.

We start by examining the size of the recent tariffs. As of the end of December 2025, the latest data used in our analyses, the actual tariff rate (12 percent) was less than half as large as the statutory rate (26 percent). Shipment lags, product and company-specific exemptions, utilization of the US-Mexico-Canada Agreement (USMCA), and evasion and heterogeneous enforcement all contribute to the large gap between statutory and actual rates. A key reason why the price impact of the tariffs remained below many forecasts made in April 2025 is that the implemented policy remains much smaller than the announced policy.

Next, we study the economic "incidence" of the tariffs, or the degree to which the tariffs' costs are borne by the United States or by the exporter on whom the tariffs are imposed. Legally, the US importer has the full responsibility to pay the tariff to the US Customs and Border Protection, typically within ten days of when imports physically enter the United States. But as economists have long taught, the economic incidence of the tariff may differ from the accounting or legal incidence. We give a theoretical overview of the key determinants of the extent to which tariff-inclusive import prices will rise in step with tariffs, a concept called tariff pass-through. Empirically, we show that pass-through is pervasively high during both the 2018–2019 and 2025 episodes. Our preferred methodology results in an estimated tariff pass-through of 81 percent during 2018–2019 and 92 percent during 2025, although the higher rate in the recent episode may reflect the shorter horizon we can study. Our estimates of very high 92 percent rate of tariff pass-through of the 2025 US import tariffs is consistent with the results in Amiti, Flanagan, Heise, and Weinstein (2026) (86–94 percent), Azzimonti and Titcomb (2026) (98 percent), Fajgelbaum and Khandelwal (2026) (90 percent), and Hinz, Lohmann, Mahlkow, and Vorwig (2026) (96 percent). Given this high pass-through to import prices and the importance of imported inputs in US manufacturing, much of the incidence of the higher US tariffs falls on US producers as well as on consumers. We discuss both below.

The recent episodes of US tariffs have led to striking changes in sourcing patterns. The share of Chinese exports in US goods imports collapsed from 22 percent at the end of 2017 to about 12 percent at the end of 2024. By December 2025, China's share was only 7 percent. Countries like Vietnam gained shares in US imports, although it is not yet clear whether these increases reflect Vietnamese value-added, as opposed to value-added in China that is now simply routed through these exporters, as discussed in Gopinath, Gourinchas, Presbitero, and Topalova (2025) and Alfaro and Chor (2025). Finally, we conclude by revisiting the typical theoretical prediction that US import tariffs should bring about a strengthening of the dollar exchange rate. After the 2018–2019 tariffs, the dollar exchange rate moved in the predicted direction. However, in 2025, the dollar instead depreciated significantly. We speculate that the deviation from that prediction may reflect forces other than the direct effect of the tariffs.

Figure 1
Statutory and Actual US Tariff Rates, Overall, 2025



Source: Gopinath and Neiman (2026).
Note: The dashed lines in Figure 1 plot versions of the statutory and actual trade-weighted tariff rates where the weight of each exporter-HTS10 changes to reflect its share of imports each month. Importers have shifted spending away from high-tariff goods, leaving these variable-weight measures below their fixed-weight counterparts. As new monthly data are released, we will post updated versions of Figure 1 on our web pages and at https://brentneiman.com/research-writings. That link shows, for example, that the values for the statutory and actual rates as of March 2026 are 21.3 percent and 9.0 percent. Both are well below their December 2025 values.

Statutory versus Actual Tariff Rates

How large were the recent trade policy shocks? US tariffs implemented during 2018–2019 roughly doubled the trade-weighted average statutory rate from 2 to 4 percent, the highest rate since the North American Free Trade Agreement (NAFTA) entered into force in 1994. The tariff increase in 2025 was far more dramatic.

The Statutory-Actual Gap in 2025

The thin solid red line in Figure 1 shows our baseline measure of the statutory trade-weighted tariff rate. To calculate this "Statutory" rate, we use data from the US Census Bureau covering the value of imports and tariffs paid on 19,000 ten-digit harmonized tariff schedule commodities (HTS10). For each month, exporter, and HTS10 category, there are observations corresponding to different districts of entry into the United States, locations of unlading, declared tariff rate provisions, and import programs. We find the maximum ratio of tariff revenues to import values across those observations and treat it as the statutory rate for that month, exporter, and HTS10. To get the aggregate plotted in the figure, we then weight

these maximum rates using their share in aggregate imports in 2024.[1] After peaking at 32.8 percent in April, the statutory rate in December sat at 25.7 percent. This is the highest level since at least the late 1930s, a time when import spending relative to GDP was less than half what it is today.

Why do we calculate statutory rates in this way, rather than constructing a real-time index based on all announcements, entries in the Federal Register, and implementation guidance from Customs and Border Protection? Our methodology avoids much of the ambiguity and discretion that such a procedure would require. We assume that the highest tariff rate that is actually paid on imports of a given HTS10 commodity shipped by a given exporter reveals the statutory rate.[2]

But when evaluating the effects of tariffs, including their timing, pass-through, and incidence, it would be incorrect to use the statutory rate. The actual tariffs that were applied in 2025—the true trade policy shocks—were much smaller. The thick solid blue line in Figure 1 shows our baseline measure of the actual trade-weighted tariff rate. To calculate this "Actual" rate, we divide the tariff revenues that the US government collects each month on imports from each exporter-HTS10 by the value of those imports. We aggregate using the same fixed trade weights from 2024 as we used for the statutory rate. The actual rate has risen far more slowly and modestly, did not experience a dramatic reversal or the volatility seen in the statutory rate, and remained at only 12.4 percent by the end of December.[3]

Drivers of the Statutory-Actual Gap

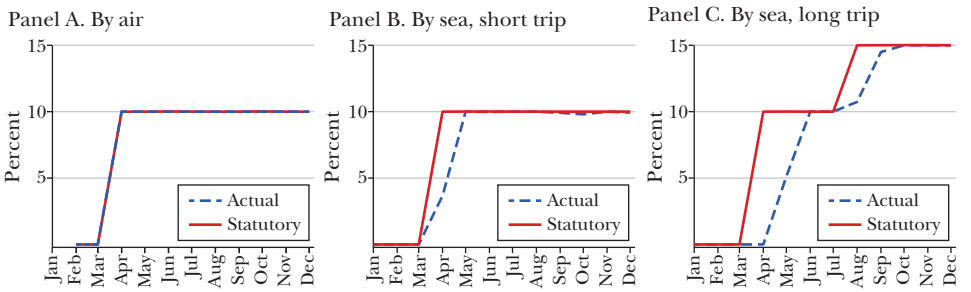
What accounts for the large gap between the statutory and actual tariff rates? Should we expect it to close over time? In principle, the statutory-actual gap reflects four factors: (1) shipment lags, (2) product and country exemptions, (3) provisions of the US-Mexico-Canada Agreement, and (4) evasion and limited enforcement.

[1]We include data through December 2025, which were published in February 2026, and note that all our results can easily be updated to more recent months using the replication package found at our Github repo at https://github.com/neiman-research/incidence_of_tariffs. We exclude a small number of observations reflecting tariffs of exactly 200 percent, which arise from duties charged on imports from third countries deemed to contain Russian-origin aluminum, as described in the Proclamation 10522 of the Federal Register on February 24, 2023. Our aggregate tariff measures are likewise not sensitive to the omission of agricultural product lines subject to specific, rather than ad valorem, tariffs (including implied rates of up to 208 percent on some out-of-quota dairy imports), as such tariffs cover a very small share of US trade and do not form part of the recent tariff waves.

[2]Our calculation of the statutory tariff rate is somewhat higher than similar measures provided by the WTO-IMF Tariff Tracker (at https://ttd.wto.org/en/reports/tariff-actions), the Trade War Tracker (at https://www.tradewartracker.com), and the Yale Budget Lab (at https://budgetlab.yale.edu/research/short-run-effects-2025-tariffs-so-far). However, all the measures show a similar evolution. Among other differences, our measure calculates statutory rates at a more disaggregated level and does not include USMCA exemptions in our definition. Our approach will understate the statutory rate if exemptions are so pervasive that no shipments of a given HTS10 and from a given exporter paid the full tariff.

[3]Feenstra and Hong (2024) find much smaller welfare costs from the 2019 tariffs than Fajgelbaum, Goldberg, Kennedy, and Khandelwal (2020b) do, and attributes the difference to the gap between actual and statutory tariffs during that episode.

Figure 2
Examples of Lags from the "On-the-Water" Exemption



Source: Gopinath and Neiman (2026).
Note: Each panel plots the statutory (solid red) and actual (dashed blue) trade-weighted tariff rates for a set of imports chosen to differ by their mode of shipment and therefore transit time. Longer transit times delay when a newly announced statutory rate binds on shipments actually received, widening the statutory-actual gap for a longer period. The gap closes fastest for air shipments (panel A) and most slowly for long ocean routes (panel C).

Shipment Lags. The tariff rate that applies to a shipment is set at the time the shipment begins its transit. This "on-the-water" exemption means that several months might pass before a newly announced statutory tariff rate applies to an import shipment that is actually received. For air shipments, which take only a few days at most, shipping lags do not matter much. For shorter ocean shipping trips, like from the United Kingdom to the east coast of the United States, the shipping lag can be a couple of months. For longer ocean shipping, like from Japan to the east coast of the United States, the lag can extend to three months. The three panels of Figure 2 show the evolution of the statutory-actual tariff rate gap using data on a small set of actual shipments with short, medium, and long transit times.

Product and Company Exemptions. Specific products or even specific companies (that commit to build manufacturing plants in the United States, say) have been offered exemptions from the tariffs. For example, various semiconductors were granted exemptions from the reciprocal tariffs announced by the United States in early April 2025. As a result, semiconductors have an actual tariff rate of only 9 percent despite a much larger statutory tariff of 22 percent. Such exemptions can also affect actual tariffs at the national level: for example, the same semiconductor exemption drives the gap between Taiwan's 27 percent statutory rate and its 9 percent actual rate.

USMCA Utilization. The US-Mexico-Canada Agreement replaced the earlier North American Free Trade Agreement in 2020. If a sufficient share of a shipment's value was added inside the three economies, rather than imported into the bloc, the importer can typically declare the goods to be "USMCA compliant" and largely

avoid tariffs. However, doing so involves documentation and record-keeping costs, so importers may only choose to use the USMCA exemption when they would otherwise face a high tariff (Hayakawa, Kim, and Yoshimi 2017). The higher US tariffs in 2025 may explain why utilization rates reported in the data jumped sharply for imports from both Canada and Mexico, from below 50 percent in 2024 to nearly 90 percent by December 2025. Any utilization of USMCA, whether at the lower or recently-higher levels, will contribute to the aggregate statutory-actual tariff rate gap.

Enforcement and Evasion. Finally, uneven enforcement or evasion of tariffs will drive a wedge between the statutory and actual rates. Some forms of evasion will not contribute to our measured gap. For example, underreporting import values or rerouting goods through third countries that face lower tariff rates will not generate the statutory-actual gap. However, if importers claim an exemption to which they are not entitled or claim a larger share of the purchase price is due to intangibles or distribution costs that are not subject to tariffs, this may contribute to the gap. For additional analyses related to the 2025 statutory-actual tariff gap, see Eck, Hoang, Mix, and Ray (2026) and Amiti, Flanagan, Heise, and Weinstein (2026).

The Statutory-Actual Gap in 2018–2019

How did the difference between statutory and actual tariff rates evolve when the United States hiked tariff rates in 2018–2019? Unlike the 2025 tariffs, which have increased statutory rates by 10 percentage points or more on 88 countries and 68 products (measured at the two-digit harmonized code level), the tariffs in 2018–2019 were more narrowly focused on a few industries and, particularly, on imports from China. There were fewer consequential product-level exemptions, no major rate reversals, and no change in the gap due to compliance with the then-binding North American Free Trade Agreement. However, exclusions from the tariffs for certain imports from China still played an important role (Chor, Grant, and Li 2025). As a result, the aggregate statutory-actual gap was small and fairly stable, growing from 0.5 to 1.7 percentage points during 2018–2019 before stabilizing at 1 percentage point by 2021, about where the gap remained until 2025.

Looking Ahead

Recent legal decisions will likely be important in shaping the statutory and actual rates—and the gap between them—moving forward. The US Supreme Court in February 2026 ruled that many of the 2025 tariffs—those enacted under the authority of the International Emergency Economic Powers Act (IEEPA) of 1977—were illegal. Shortly after the Supreme Court ruling, the administration enacted tariffs based on Section 122 of the Trade Act of 1974. However, the US Court for International Trade ruled in May 2026 that the Section 122 tariffs were also unlawful.

If, despite these legal rulings, tariffs remain in place at levels comparable to what they were at the end of 2025, should we expect the actual rate to "catch up"

with the statutory rate? Probably not. While effects of the "on-the-water" exemption will fade, USMCA utilization will likely remain high, evasion efforts will become more sophisticated, and the announcement of new exemptions may continue, such as those in late 2025 applying to bananas and coffee.

Increases in statutory tariff rates are easy to announce and quick to disseminate to the public. However, the actual tariff rates paid by importers are often substantially lower due to exemptions and may only bind after a long delay. The rest of the paper, therefore, focuses only on the actual tariff rate, not the announced or statutory rate. We assess the incidence and pass-through of the tariff that has actually been implemented.

Determinants of Pass-Through

In this section, we describe the theoretical determinants of tariff pass-through and compare them to determinants of exchange rate pass-through. We explain why the two pass-through rates need not be the same. In the next section, we will then offer estimates of the rates of pass-through after the 2018–2019 and 2025 tariffs were imposed. Consider the case of a foreign exporter choosing flexibly the (log) dollar price, p_F, at which to sell to the US market when facing a tariff rate of τ and an exchange rate (in logs) of e, where a higher value of e implies a weaker US dollar. The effect of changes in tariffs and exchange rates on the tariff-inclusive dollar import price is given by:[4]

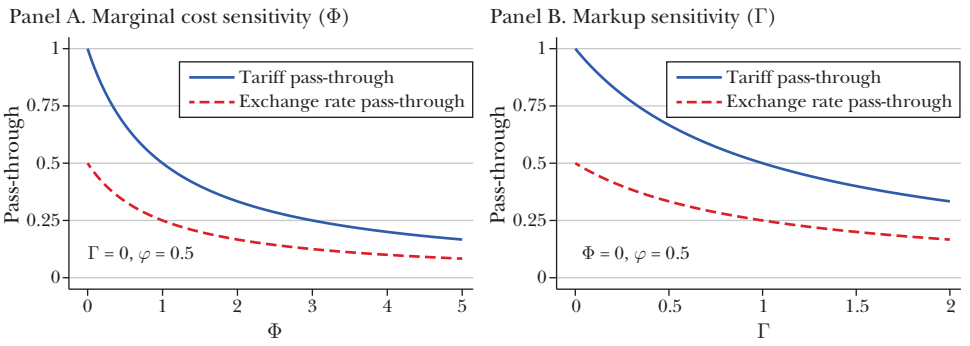
$$\Delta p_F + \Delta\tau = \underbrace{\frac{1}{1+\Gamma+\Phi}}_{\text{Tariff pass-through}}\Delta\tau + \underbrace{\frac{\phi}{1+\Gamma+\Phi}}_{\text{Exchange rate pass-through}}\Delta e,$$

where the operator Δ represents a very small change. This includes three parameters (Φ, Γ, and ϕ) that we describe in turn.

The first parameter, Φ, captures the sensitivity of the marginal cost of the exporter to a change in the price it sets. If the exporter's production function displays constant returns to scale, then its marginal cost is insensitive to the price it sets and $\Phi = 0$. If the exporter's production function displays decreasing returns to scale, then its marginal cost falls if it sets higher prices (and sells lower quantities) and $\Phi > 0$. Specifically, the parameter Φ is the product of the elasticity of the firm's marginal cost to quantity and the elasticity of demand to the price that the firm sets. If the marginal cost and demand are more elastic, then the pass-through of the tariff to export prices will be lower, and therefore to the tariff-inclusive import

[4]See the Supplemental Appendix for additional details and derivations of all the results in this section, which closely follows Gopinath, Itskhoki, and Rigobon (2010) and Burstein and Gopinath (2014), extended to the case of tariffs. Here, we focus on the "direct" effect of tariffs, by which we mean setting aside the indirect effect on the firm's prices through changes in the industry price level. If the firm is small, then the indirect effect is negligible. Our discussion here also assumes all domestic costs in the exporter country, such as wages, are unchanged.

Figure 3
Theoretical Pass-Through to Tariff-Inclusive Price



Source: Gopinath and Neiman (2026).
Note: Both panels plot tariff pass-through (solid blue) and exchange rate pass-through (dashed red) to the tariff-inclusive import price, assuming a domestic input cost share of $\phi = 0.5$. Panel A holds the markup sensitivity fixed ($\Gamma = 0$) and varies the sensitivity of marginal cost to the firm's relative price (Φ); panel B holds marginal cost sensitivity fixed ($\Phi = 0$) and varies markup sensitivity (Γ).

higher. Conversely, the more inelastic the exporter's supply is relative to demand, the greater the burden borne by the exporter.

The second parameter, Γ, captures the sensitivity of the exporter's markup to the price it sets. For example, if the elasticity of demand that the exporter faces is constant, then it will price at a constant markup over marginal cost and $\Gamma = 0$. If, however, the elasticity of demand increases when the firm's price increases relative to its competitors, then the firm's optimal markup increases with its market share and $\Gamma > 0$. If an exporter with $\Gamma > 0$ faces a new tariff but does not reduce its price, it will lose market share, leaving the existing markup too high. Profit-maximization would instead call for the firm to reduce its markup and, therefore, its export price.[5] As shown in Amiti, Itskhoki, and Konings (2019), this channel is particularly prominent for firms with larger market shares as compared to small firms, consistent with theory.

The impact of these forces on the pass-through of tariffs is represented in the solid blue lines in Figure 3. Panel A of Figure 3 assumes a constant elasticity of demand ($\Gamma = 0$) but considers variation in the sensitivity of marginal costs to changes in a firm's relative price. When the exporter's marginal cost is constant, $\Phi = 0$ and pass-through is 100 percent, the value where the solid line intersects the *y*-axis. Moving to the right of the plot, pass-through drops as the marginal cost sensitivity increases. Similarly, panel B of Figure 3 assumes a constant marginal cost but considers variation in the sensitivity of markups to changes in a firm's

[5]The magnitude of this channel depends on the elasticity of the elasticity of demand, referred to as the super-elasticity of demand, as discussed in Kimball (1995) and Klenow and Willis (2016).

relative price. When the exporter's markup is constant, $\Gamma = 0$ and pass-through is 100 percent. Pass-through drops as the sensitivity of markups grows. Note that when the exporter's marginal cost and the elasticity of demand are both constant, $\Gamma = \Phi = 0$, pass-through is 100 percent, and the full price incidence of the tariff is borne by the importer. Even with full pass-through, the exporter will still be worse off, of course. Its per-unit markup will not change, but it will sell a smaller quantity.

How does tariff pass-through compare to exchange rate pass-through? As shown in the equation, the factors Γ and Φ influence tariff and exchange rate pass-through in identical ways. However, an additional factor capturing the share of domestic inputs (such as labor) in production costs, $\phi \in [0,1]$, also influences exchange rate pass-through. Most exporters rely on imported inputs, often rigidly priced in US dollars, for production. Spending on foreign inputs drives ϕ down, reducing the fraction of a firm's marginal cost that is sensitive to the exchange rate and, consequently, reducing exchange rate pass-through. There is extensive empirical evidence for this channel in the literature, including evidence for Belgian firms in Amiti, Itskhoki, and Konings (2014).

Both panels of Figure 3 depict the case when $\phi = 0.5$. As a result, the values along the dashed red line that plots exchange rate pass-through are half the corresponding values on the solid blue line plotting tariff pass-through. Our discussion has focused on a static environment, but expectations can also matter if one considers a richer dynamic framework.[6]

To summarize, tariff pass-through can vary depending on the curvature of the production function and on the elasticity and super-elasticity of demand. If the marginal cost of production and the elasticity of demand are constant, then exporting firms fully pass-through the tariff. Even when tariff pass-through is high, exchange rate pass-through can be lower when production relies on imported inputs or when the exchange rate changes are perceived as being more transitory than a random walk.

Pass-Through to Import Prices

In this section, we measure pass-through of tariffs and exchange rates to prices paid by US importers. We leave the discussion on the impact of tariffs on exchange rates for later.

Tariff Pass-Through Using BLS Import Price Indices, 2025

To observe changes in the price of US imports, we start with indices constructed by the International Pricing Program of the US Bureau of Labor Statistics (BLS). These indices measure monthly price changes for imported goods whose quality or

[6]For example, as discussed in Supplemental Appendix Section A.2, if changes in tariffs are expected to be longer-lasting than changes in nominal exchange rates, then tariff pass-through may exceed exchange rate pass-through even when $\phi = 1$.

characteristics are fixed over time. When needed, hedonic and other techniques are used to create quality-adjusted prices. Observing the behavior of import prices using the BLS price indices, rather than unit values in the Census trade data, ensures that price changes are isolated from shifts in the composition of purchases to different goods. Therefore, our initial analysis of pass-through focuses on the set of countries, economies, and sectors where the BLS provides indices.

To get a sense for the raw data that we will use in our calculations, we use the twelve-month change through December 2025 in the (exclusive of tariffs, or "ex-tariff") price of US imports against the increase in the tariff rate applied on that exporter over that same period for all exporters reported by the BLS.[7] Of the seven exporters or exporting regions that experienced tariff increases of 5 percentage points or more, only shipments from China exhibited a decline, equal to 3.1 percent, in their ex-tariff price. The price of imports from the others were essentially unchanged or rose by roughly 2 percent or less. Using these observations, we calculate the tariff-inclusive pass-through to import prices for these exporters as:

$$\textit{Pass-Through to Import Prices} = \frac{\textit{Ex-Tariff Price Change} + \textit{Tariff Change}}{\textit{Tariff Change}}.$$

Panel A of Table 1 lists the results for each source of US imports. The first column shows the actual tariff. The final column shows that pass-through rates to the "raw" import price indices for these 7 exporters or exporting regions were all high or in excess of 100 percent.

One concern with this simple approach is that the price of imports from each exporter may have been experiencing different trends prior to the imposition of the 2025 tariffs. However, when we plot the detrended change in each of the ex-tariff import price indices, which we then use for our preferred pass-through measures in the BLS data, little changes.[8] As shown in the second column of panel A of Table 1, the pass-through rate calculated using this detrended method drops meaningfully for the European Union, Latin America, and the United Kingdom, but the resulting mean and median pass-through rates remain high at roughly 90 percent. Results are qualitatively similar if we instead focus on the acceleration of the price change

[7] Supplemental Appendix Figure A1a presents a plot of these data.

[8] The calculations are available in Supplemental Appendix Figure A1b. We detrend each monthly price index by fitting a LOESS curve to time using data only through the last pre-shock month, then subtracting the fitted values from the original series. LOESS is a locally weighted, nonparametric regression that fits low-degree polynomials to nearby observations, weighting closer months more heavily. For more information, see Cleveland, Cleveland, McRae, and Terpenning (1990). We set the smoothing span to 10 percent, so each point's trend reflects the nearest 10 percent of observations in time. For example, this equals three months where we have 30 months of pre-tariff price data. Limiting the fit to pre-shock data ensures the estimated trend is not affected by the shock, so post-shock residuals capture deviations from the pre-shock path.

Table 1
Tariff Pass-Through Summary, Countries and Sectors

Country/region	Tariff change (pp)	Detrended price pass-through	Price acceleration pass-through	Raw price pass-through
Panel A. Countries and regions				
China	18.5	0.87	0.87	0.83
Japan	12.7	1.10	1.13	1.11
ASEAN	12.4	1.20	1.15	1.18
European Union	8.0	0.83	0.79	1.02
Taiwan	7.8	0.98	1.07	1.20
Latin America	6.0	0.73	0.53	1.09
United Kingdom	5.8	0.70	0.59	0.88
Mean	10.2	0.92	0.88	1.04
Median	8.0	0.87	0.87	1.09
Sector (NAICS)	**Tariff change (pp)**	**Detrended price pass-through**	**Price acceleration pass-through**	**Raw price pass-through**
Panel B. Sectors				
332—Fabricated Metals	24.8	0.83	0.86	0.84
314—Textile Products	22.7	–	0.73	0.72
33992—Sporting & Athletic Goods	20.4	0.94	0.91	0.96
3331—Ag., Constr. & Mining Machinery	20.1	1.11	1.08	1.12
33991—Jewelry & Silverware	19.6	1.62	1.65	1.81
33993—Dolls, Toys & Games	19.1	1.06	1.03	1.02
3336—Engines, Turbines & Power Trans.	18.4	1.16	1.21	1.15
315—Apparel	18.1	0.84	0.84	0.84
3339—Other General-Purpose Machinery	17.7	1.09	1.08	1.08
331—Primary Metals	17.6	1.59	1.66	2.34
313—Textiles	16.4	–	0.68	0.82
33699—Other Transportation Equip.	16.3	–	1.06	0.99
112—Animal & Aquaculture	15.5	–	0.35	1.08
33634—MV Brake Systems	14.9	0.78	0.83	0.76
335—Electrical Equip. & Components	14.9	1.31	1.31	1.17
326—Plastics & Rubber	14.0	0.74	0.73	0.75
3335—Metalworking Machinery	13.9	1.49	1.05	1.19
327—Nonmetallic Minerals	13.8	0.73	0.72	0.75
33635—MV Transmission & Power Train	13.7	1.02	1.00	1.10
3333—Commercial & Service Machinery	13.3	1.06	1.00	0.98
33611—Autos & Light Vehicles	13.2	0.63	0.61	0.76
3343—Audio & Video Equip.	12.3	0.91	0.90	0.97
33639—MV Other Parts	12.2	0.93	0.87	1.05
33633—MV Steering & Suspension	11.7	–	0.94	1.03
33631—MV Gasoline Engines & Parts	10.6	1.06	0.90	1.22
33632—MV Electrical & Electronic Equip.	9.6	−0.17	−0.27	0.76
33911—Medical Equip. & Supplies	9.4	–	1.05	1.18
322—Paper	8.7	0.32	0.43	0.23
3252—Resins & Syn. Rubber/Fibers	7.6	–	0.04	−0.14
312—Beverage & Tobacco	7.5	−1.33	−1.35	−1.15
3251—Basic Chemicals	7.1	0.19	0.44	−0.19
311—Food	7.0	0.64	0.16	1.44
3344—Semiconductors & Elec. Components	6.7	1.15	1.16	1.16
3342—Communications Equip.	6.0	0.15	0.18	0.03
Mean	14.0	0.81	0.76	0.88
Median	13.9	0.93	0.88	0.98

Source: Gopinath and Nieman (2026).

during the twelve months ending December 2025 relative to the twelve months ending December 2024, as reported in the third column of panel A of Table 1.

The measures of pass-through in Table 1 are admittedly crude. They look only at twelve-month variation, pool goods and countries with heterogeneous tariff rates, and associate all price increases, or above-trend increases, with the tariffs. That said, we find no evidence in the country- and region-level BLS import price indices of ex-tariff price declines that largely offset the imposed tariffs.

We can also use the BLS data to calculate pass-through at the sector level. To balance the desire to study disaggregated sectors while maximizing coverage of the US import basket, we assemble a set of 44 non-overlapping industrial classification (NAICS) codes ranging from the three- to five-digit levels where BLS price indices are available.[9] The pass-through rates for those sectors with tariff increases of at least 5 percentage points are shown in panel B of Table 1. Detrended pass-through rates for sectors with at least a 15 percentage point increase in tariffs (and with enough data for our detrending algorithm to work) were all above 83 percent. Across all sectors, detrended pass-through had a mean value of 81 percent and a median value of 93 percent.

What might explain the heterogeneity in these pass-through rates? As discussed earlier, differences in the properties of demand or of the production function will matter. The properties may reflect the extent to which US imports in a sector come from a small and concentrated set of export partners and the extent to which US producers are themselves active in the sector. However, such characteristics do not appear to be clearly related to pass-through in the data.

For example, we compared each sector's pass-through rate with its Herfindahl-Hirschman concentration index (HHI) of supplier countries, but did not find a close relationship. The two sectors with highest HHIs are "Doll, Toy, and Game Manufacturing," where nearly 80 percent of imports in 2024 were from China, and "Motor Vehicle Seating and Interior Trim Manufacturing," where nearly 70 percent of imports were from Mexico. Those two highly concentrated sectors, however, have pass-through rates of essentially 100 percent, no different from the large mass of sectors with low concentration index values. Similarly, we do not find evidence that sectors in which the United States exports a lot—and therefore, presumably, has capacity to produce substitutes for the imported goods—had systematically different tariff pass-through rates compared with sectors that exported very little.

In short, whether looking across exporters or sectors, pass-through rates are heterogeneous but typically high. Through December 2025, the price impact from recent US import tariffs, in general, is borne by the US importer.

[9]We start by selecting all NAICS three-digit price indices with sufficient historical coverage in the BLS data. Next, we consider the combination of non-overlapping four- and five-digit codes that correspond to each three-digit code and are also available in the BLS data. If those more disaggregated sectors account for at least 75 percent of the total value of imports in 2024 in the three-digit code, we use them instead of the three-digit code. If not, we disregard them and only use the three-digit sectors. These sectors are listed in Supplemental Appendix Table A1 and collectively represent 83 percent of the value of US merchandise imports in 2024.

Table 2
Tariff Impacts on Import Values and Quantities: OLS

	log change tariff-inclusive import prices $\Delta\ln(\tilde{p}_{ijt}(1+\tau_{ijt}))$ (1)	log change tariff-inclusive import prices $\Delta\ln(\tilde{p}_{ijt}(1+\tau_{ijt}))$ (2)	log change import quantities $\Delta\ln(\tilde{q}_{ijt})$ (3)	log change import values $\Delta\ln(v_{ijt})$ (4)
Panel A. 2018–2019				
$\Delta\ln(1+\tau_{ijt})$	0.801	0.807	−2.504	−2.814
	(0.035)	(0.035)	(0.110)	(0.091)
$\Delta\ln(e_{jt})$		0.065	−0.228	−0.213
		(0.017)	(0.056)	(0.046)
Observations	64,150	63,791	63,791	101,796
R^2	0.154	0.155	0.169	0.132
Panel B. 2025				
$\Delta\ln(1+\tau_{ijt})$	0.891	0.916	−0.927	−1.127
	(0.019)	(0.020)	(0.056)	(0.052)
$\Delta\ln(e_{jt})$		0.099	−0.402	−0.346
		(0.023)	(0.070)	(0.064)
Observations	87,392	85,147	85,147	114,494
R^2	0.181	0.184	0.159	0.129

Source: Gopinath and Nieman (2026).

Tariff Pass-Through Using Census Unit Values

The country and sector data from the US Bureau of Labor Statistics show generally high pass-through rates to import prices, but offer limited variation to identify the drivers of higher or lower rates. In this subsection, we therefore follow Amiti, Redding, and Weinstein (2019) in studying tariff pass-through using data on unit values from the Census trade data.

Unit values ($\tilde{p}_{ijt}$) can be computed by dividing the value of all shipments for a given HTS10 (i) from a given exporter (j) in a given month (t), v_{ijt}, by the reported units ($\tilde{q}_{ijt}$) shipped:

$$\tilde{p}_{ijt} = \frac{v_{ijt}}{\tilde{q}_{ijt}}.$$

We use tildes to acknowledge the well-known limitations that changes in unit values and quantities may be affected in trade data by changes in composition of goods and heterogeneity in the notion of units used, even within a given HTS10. If the impact of those limitations is minimal, the percent changes in import unit values and units should approximate the percent change in product prices and quantities, and we use this as the basis for our regression framework below.

Table 2 shows results from regressing the change in tariff-inclusive unit values, units, and ex-tariff import values on changes in the relevant actual tariff rate and

bilateral exchange rate. We start by studying the 2018–2019 tariffs in panel A, where changes are measured from December 2017 to December 2019.[10] This window is long enough to capture multiple waves of tariffs during the 2018–2019 period and to allow for a lagged response.

For the time period 2018–2019, the coefficient in column 1 in panel A of Table 2 equals 0.801 and is tightly estimated. This corresponds to a tariff-inclusive pass-through rate of 80 percent, somewhat lower than estimates of that same episode found in Cavallo, Gopinath, Neiman, and Tang (2021), Fajgelbaum, Goldberg, Kennedy, and Khandelwal (2020a), and Amiti, Redding, and Weinstein (2019), but still high. Column 2 reports results from a specification that adds the exchange rate, where an increase in the variable corresponds to a depreciation of the US dollar. The tariff pass-through rate implied by this specification remains near 80 percent, and exchange rate pass-through is estimated to be low, at 6.5 percent. This estimate of exchange rate pass-through is qualitatively similar to, but below, the 20 percent rate estimated by Cavallo, Gopinath, Neiman, and Tang (2021) from more disaggregated, transaction-level, data during that same period. The coefficient in column 4 suggests import values dropped by nearly three log points per log point increase in tariffs, consistent with a trade elasticity of about four. Results (for both panels) are qualitatively similar when we consider alternate specifications including weighted least squares or conditioning on proxies for prices in the exporting country.

Panel B of Table 2 shows results from equivalent regression specifications as shown in panel A, but with changes measured from December 2024 to December 2025. Tariff pass-through in these regressions is somewhat higher, with the tightly estimated coefficient of 0.916 in the second column suggesting a tariff-inclusive pass-through rate of 92 percent. Exchange rate pass-through is also slightly higher, at about 10 percent. These results are broadly similar with those from the 2018–2019 episode, though the shorter horizon (particularly because most tariff changes occurred only in April) suggests we may only be capturing part of the price and quantity response.

How Have Imports Responded?

How did goods imports respond to these tariff episodes? During 2018–2020, when the statutory and actual tariff rates roughly doubled, imports exhibited a clear decline from roughly 12 percent to just above 10 percent of GDP. Imports subsequently recovered, though, and exceeded 11 percent of GDP on the eve of the much-larger tariffs of 2025. Goods imports and the goods trade deficit surged

[10] Observations are by HTS10 import category and by country. All columns include product fixed effects. Standard errors are clustered at the HS8 level. Venezuela is excluded. For regressions where the dependent variables are unit values or quantities, we exclude observations where the year-on-year change exceeds a factor of three.

upward in December 2024 and in the first few months of 2025, presumably in anticipation of the large tariffs put in place in April. It is difficult to know the extent to which this "front-running," as opposed to the subsequently imposed tariffs, drove imports downward in April, May, and June. However, goods imports and the goods deficit finished 2025 at levels slightly below their averages during 2024. It would be premature to draw strong conclusions about the effect of the 2025 tariffs on total US imports of goods or the US trade deficit.

By contrast, it seems clear that tariffs, both in 2018–2019 and in 2025, have contributed to the falling share of US imports coming directly from China. In the years immediately prior to the 2018 tariffs, China's share of US imports fluctuated around 22 percent. The imposition of tariffs on China in 2018 kicked off a dramatic secular decline, nearly halving China's share in US imports to 12.5 percent by the end of 2024. Interestingly, this decline continued through the 2021–2024 period, despite stable tariff rates during those years. Coinciding with the acute spike in both tariff rates during 2025, China's share of US imports has collapsed further, reaching close to 7 percent at the end of 2025.[11] Chinese exports are potentially being rerouted through "connector countries" like Vietnam, in addition to Chinese production itself shifting to countries like Vietnam and Mexico (Gopinath, Gourinchas, Presbitero, and Topalova 2025).

In 2025, China has been subject to the largest increase in US tariffs, and China's exports to the United States have dropped most among large US trade partners. Beyond that, however, the correlation between tariff increases and import spending among other countries is limited. India and Vietnam, for example, have substantially grown their exports to the United States, but this likely represents the fact that, compared to other trade partners, their production is more substitutable for goods that the United States historically sourced from China.

Pass-Through to Producer Prices

Tariff-inclusive pass-through to US import prices is high. When a 10 percent tariff is imposed on an imported good, US importers appear to pay 8–10 percent more, including the tariff, for that good. Our analyses reveal this whether looking at aggregated countries and sectors or looking at shipments disaggregated to the exporter-HTS10 level. This was true both in 2018–2019 and in 2025, and is consistent with the large shift in US import shares—particularly the shift away from Chinese exports. What does this mean for US producers? Unlike in the past, final goods today represent less than half of all trade. Most US goods imports are intermediate inputs or capital goods used by US firms to produce. In this section, we

[11] If this collapse of trade from China resulted in the cessation of imports in a given HTS10, it might result in missing values that would impair our calculation of statutory and actual tariffs. We have verified, however, that the scale of entry and exit of exporter-HTS10 combinations during this period, does not appear unusual relative to prior periods.

focus on the implications of the tariffs for the production costs of US firms, particularly manufacturers.

Tariffs, Intermediates, and the Input-Output Structure

The US Bureau of Economic Analysis (BEA) produces a "Trade in Value Added" (TIVA) dataset that links input-output tables with trade data to show the import content of US production, by industry and source region. The most recent TIVA data are from 2023 and cover the entire US economy, disaggregated into 45 agriculture, manufacturing, and mining sectors and 93 service sectors. We combine this information on the global sourcing structure of US industry, corresponding measures of actual tariffs, and our findings of near-complete pass-through to US import prices in 2025 to calculate the extent to which the tariffs may affect the cost of US producers.[12]

For example, the data suggest that imported inputs accounted for 23 percent of the value of production of US-produced heavy-duty trucks. About one-quarter of that imported input spending was on vehicle parts from China, Japan, countries in the "rest of Asia," and Europe according to the BEA's allocation of imports across origins. We estimate that tariffs on vehicle parts from China increased in 2025 by about 29 percentage points, and tariffs on parts from other regions increased by about 17 percentage points. Together with the other three-quarters of input spending, we multiply the increase in trade-weighted average tariff by region and imported input into the US heavy-duty truck sector by their share in total production costs, which we call the increase in "production tariff" of 3.2 percent. In other words, given the pattern of global sourcing, our calculation suggests that increased tariffs impact the industry similarly to a hypothetical new tax on total production costs of 3.2 percent.

Production Tariffs

Consider the five production sectors most affected by the tariffs: "Heavy-Duty Trucks" has a 3.2 increase in its production tariff and "Construction Machinery" has a 2.8 percentage point increase because imported steel accounts for a large share of their production costs; "Motor Vehicles" and "Motor Vehicle Parts" each have increases of 2.3 percentage points due to reliance on both steel and auto inputs; and "Agricultural Implements" has a 2.2 percent production tariff increase because 13 percent of its production costs involve the use of other imported agricultural

[12]The TIVA methodology is found at https://www.bea.gov/system/files/methodologies/Technical-document-methodology-for-preparing-single-country-trade-in-value-added-statistics.pdf and https://apps.bea.gov/national/TiVA/bea-TiVA-table-builder-user-guide.pdf. The BEA allocates imported commodities to industries using the assumption that industries source imports in proportion to each origin's share of total US imports of that commodity. See Johnson and Noguera (2012) for work that was foundational to the development of the TIVA data, and Minton and Somale (2025) for an example of an approach related to ours that links tariffs to personal consumption expenditure index inflation.

implements, many of which come from China.[13] Aggregating across all subsectors, we calculate that the production tariff increased by 0.94, 0.39, and 0.12 percentage points, respectively, in US manufacturing, agriculture, and services. If we were to express these as equivalent to a tax on intermediate input spending rather than total production costs, these increases would be 1.64, 0.69, and 0.31 percentage points.

Is there evidence that these increases in production tariffs on US industries have led to increases in the price of US production, or "producer price index" (PPI) inflation? We compared twelve-month changes in PPI inflation by industry with the increase in industry-specific production tariff. Service sectors, which rely far less on intermediate inputs—both domestically produced and imported—experienced only small increases in production tariffs, always less than 0.5 percentage points and typically closer to zero. (The sole service sector exception with a higher tariff was wholesale pharmaceutical distributors.)

However, manufacturing sectors experienced changes in their production tariffs ranging from near-zero to over 3 percentage points. We use the same method as earlier to detrend sectoral changes in the Production Price Index and find evidence mildly suggestive that tariffs on these industries have pass-through into the prices charged by US manufacturers. We divide the manufacturing sectors into three groupings based on the scale of increases in their production tariffs: (1) those with the smallest increases (less than 1 percentage point), (2) those with intermediate increases (between 1 and 2 percentage points), and (3) those with largest increases (greater than 2 percentage points). First, of the 27 manufacturing sectors with small increases in production tariffs, 15 (or 56 percent) exhibited above-trend price increases. Next, of the 17 manufacturing sectors with intermediate increases in production tariffs, 13 (or 76 percent) exhibited above-trend increases. Finally, of the eight manufacturing sectors with large increases in production tariffs, six (or 75 percent) exhibited above-trend increases.

Future work using detailed firm-level data will be needed to trace the implications of import tariffs through to producer prices more convincingly and precisely, but even these crude aggregate analyses suggest that Production Price Index inflation is more likely to be above trend in US manufacturing industries whose production costs appear to be more impacted by the recent increase in import tariffs.

Pass-Through to Consumer Prices

To what extent have the import tariffs passed through to consumer prices, or Consumer Price Index (CPI) inflation? The CPI indices are based on products, rather than at the industry level, so the same approach we used to analyze producer

[13] For comparison, Supplemental Appendix Table A2 lists the ten sectors with largest increases in production tariffs.

price indices cannot be used to answer this question, at least not in a simple and straightforward way.[14]

Instead, for purposes of illustration, we considered six products in the Consumer Price Index that are imported as final goods and for which there are few or no domestically produced substitutes: bananas, children's footwear, coffee, shelf-stable fish and seafood, televisions, and toys. Some appear to show price increases relative to trend that coincide with the imposition of tariffs (though coffee's price rise appears to start a bit earlier, in late 2024). In fact, presumably due to the salience of these goods and the uptick in their price growth, bananas and coffee were exempted from tariffs by an executive order issued in November 2025. Other sectors, like Children's Footwear, do not exhibit any obvious deviations from their pre-tariff price trends.[15]

In a study of the pass-through of the 2018 tariffs, Cavallo, Gopinath, Neiman, and Tang (2021) used far more detailed data on prices at big-box chains to demonstrate low retail pass-through, at least in the first year after the tariffs were applied. The paper offered evidence that substitution away from China as a supplier and large inventories accumulated ahead of the tariffs likely contributed to the ability to avoid or forestall price increases. Cavallo, Llamas, and Vazquez (2025) use related, but enriched, data to study the 2025 tariffs and find evidence that pass-through appears to be swift. They estimate a pass-through rate of up to 24 percent to retail within a few months, and demonstrate how pricing implications from the tariffs also spill over to domestically produced goods that now face less stiff competition. Through the end of September, Cavallo, Llamas, and Vazquez (2025) find that tariffs contributed about 0.76 percentage points to the all-items Consumer Price Index.

Tariffs and Exchange Rates

In the above sections, our theoretical and empirical results address the pricing implications of both tariff and exchange rate shocks, regardless of whether tariffs or exchange rates moved in the same or opposite directions. In this section, we focus on why tariff changes may lead directly to exchange rate changes.[16]

Should Tariffs Cause the Currency to Appreciate? What Does Theory Say?

Consider the case where tariffs do not change the trade deficit. This case is useful because of its theoretical appeal. Unless extreme, permanent tariffs will only minimally alter the difference between national savings and national investment and therefore the trade balance, as discussed in International Monetary Fund

[14] Moreover, capturing the true pass-through to consumers may require a broader notion of how consumers pay for bundles of goods and services. Hankins, Momeni, and Sovich (2026), for example, demonstrate that the tariffs placed on metals in 2018 also resulted in higher financing costs for purchases of affected goods.

[15] For figures showing the increase in these six products according to Consumer Price Index Data with the increase in tariffs on these products in 2025, see Supplemental Appendix Figure A2.

[16] Details and derivations of these arguments are available in the Supplemental Appendix.

(2026). Consistent with theory, US trade deficits on an annual basis did not change much in 2025 (3.0 percent of GDP) relative to 2024 (3.1 percent).

Further, we assume that monetary policy successfully targets a fixed aggregate price level. We also assume the consumption level remains unaffected, which is reasonable when the importing country is, like the US economy, relatively closed. In this case, the decline in the exclusive-of-tariffs dollar value of imports from the increase in tariffs must be matched with an equal decline in the dollar value of exports. This decline in exports is caused by an appreciation in the dollar. Because, for a given tariff, the decline in imports will be a function of the extent of pass-through, the scale of the required dollar appreciation will also relate to pass-through.[17]

For the US economy, a US dollar appreciation in response to higher tariffs reduces the dollar price of imports, in this way partially offsetting the increase in the dollar price from the higher tariff rate, and consequently moderates the decline in imports and in exports alongside an unchanged trade balance. The scale of appreciation of the exchange rate (e) is therefore a function of the same three parameters discussed earlier that impact pass-through (Φ, Γ, and ϕ).

For simplicity, and consistent with the high pass-through rates we found empirically in the earlier discussion, we assume a constant marginal cost and constant elasticity of demand, such that $\Gamma = \Phi = 0$. When $\phi = 0$, all of the costs of the exporter are unchanged in dollars and, consequently, exchange rate pass-through equals 0. In this case, the dollar appreciates one-to-one with the size of the tariff increase. As ϕ increases, pass-through increases, and the extent of dollar appreciation declines.[18]

Have Recent Exchange Rate Movements Tracked Tariffs?

When import tariffs were imposed on China in 2018–2019, actual exchange rate changes were larger but directionally consistent with these theoretical predictions. As we discussed earlier, exchange rate pass-through into US import prices was close to zero, so we might expect the dollar to appreciate by an amount comparable to the actual tariff increase. Indeed, from early 2018 to late 2020, the actual tariff on US imports from China increased by about 8 percentage points, which was roughly the same magnitude as the US dollar's appreciation relative to the Chinese renminbi during that same period.

In 2025, however, the dollar weakened, not strengthened. The tariff rate on Chinese imports surged, yet the dollar clearly depreciated against the renminbi. Similarly, aggregate tariffs have risen sharply, and the trade-weighted dollar has

[17] Specifically, in the case of flexible prices, it can be shown that:

$$\frac{\Delta e}{\Delta \tau} = \frac{-(\sigma + \Gamma + \Phi)}{\phi(\sigma - 1) + (\sigma + \Gamma + \Phi)} < 0,$$

where $\sigma > 1$ is the common elasticity of demand for both exports and imports.

[18] Supplemental Appendix Figure A3 depicts the dollar appreciation that would follow a 1 percentage point increase in tariffs as a function of the share of the exporter's costs that are denominated in the exporter's currency, ϕ.

significantly depreciated. There is only a weak relationship between the cross-country distribution of bilateral tariffs and the cross-country pattern of bilateral exchange rate changes.

Unlike 2018–2019, the recent dollar movement has defied theory by depreciating instead of appreciating. One possibility is that our focus should shift to richer theories that allow for import tariffs to cause the importer's currency to weaken directly, such as the wave of recent work including Auclert, Rognlie, and Straub (2025), Costinot and Werning (2025), Itskhoki and Mukhin (2025), and Werning, Lorenzoni, and Guerrieri (2025). More likely, we believe, is that other policies and macroeconomic forces have counterbalanced the effects of higher tariffs on the US dollar exchange rate and pushed more forcefully in the other direction. For example, when "Liberation Day" tariffs were announced in April 2025, it led to a large surge in uncertainty and to expectations that the US economy would slow, prompting the Fed to cut interest rates faster than it might otherwise have done—leading to a dollar depreciation. Consistent with and potentially amplifying this dynamic, reports suggest foreign investors in 2025 started hedging their dollar exposure to a greater extent than in the past (Shin, Wooldridge, and Xia 2025).

Conclusion

At the end of 2025, the actual tariff rates on US imports were not nearly as large as policy announcements suggested, but they were still historically large and have led to a reshaping of US trade patterns. Pass-through to import prices is high, China's share of US imports has collapsed, and US manufacturers face higher input costs. Since then, legal rulings against the tariffs imposed under the International Emergency Economic Powers Act of 1977 and under the Section 122 of the Trade Act of 1974 have shifted expectations about the future US tariff regime. But beyond the legal issues, important economic questions remain. If tariff rates stay roughly where they are now, will the statutory-actual gap close, expand, or persist? Will pass-through into tariff-inclusive import prices remain high as more time passes? How much of the shift in US import patterns simply reflects transshipment of Chinese value-added? How will US manufacturing respond to the production tariffs that we calculate? What might be behind the depreciation in 2025 of the US dollar, which is at odds with many standard models of the effects of tariffs? We hope our work offers helpful context in evaluating the initial implications of the 2025 US import tariffs and look forward to new evidence and methods that will evolve to address these questions.

■ *The authors thank Ben Fraser for excellent research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect the position of any institution with which they are affiliated. Any errors or omissions are their own.*

References

Alfaro, Laura, and Davin Chor. 2025. "An Anatomy of the Great Reallocation in US Supply Chain Trade." NBER Working Paper 34490.

Amiti, Mary, Chris Flanagan, Sebastian Heise, and David E. Weinstein. 2026. "Who Is Paying for the 2025 U.S. Tariffs?" Liberty Street Economics, February 12. https://doi.org/10.59576/lse.20260212.

Amiti, Mary, Oleg Itskhoki, and Jozef Konings. 2014. "Importers, Exporters, and Exchange Rate Disconnect." *American Economic Review* 104 (7): 1942–78.

Amiti, Mary, Oleg Itskhoki, and Jozef Konings. 2019. "International Shocks, Variable Markups, and Domestic Prices." *Review of Economic Studies* 86 (6): 2356–402.

Amiti, Mary, Stephen J. Redding, and David E. Weinstein. 2019. "The Impact of the 2018 Tariffs on Prices and Welfare." *Journal of Economic Perspectives* 33 (4): 187–210.

Auclert, Adrien, Matthew Rognlie, and Ludwig Straub. 2025. "The Macroeconomics of Tariff Shocks." NBER Working Paper 33726.

Azzimonti, Marina, and Jacob Titcomb. 2026. "U.S. Import Tariffs in 2025: Realized Tariff Rates, Import Prices, and Local Labor-Market Effects." Federal Reserve Bank of Richmond Working Paper 26-08.

Burstein, Ariel, and Gita Gopinath. 2014. "International Prices and Exchange Rates." In *Handbook of International Economics*, Vol. 4, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 391–451. North-Holland.

Cavallo, Alberto, Gita Gopinath, Brent Neiman, and Jenny Tang. 2021. "Tariff Pass-Through at the Border and at the Store: Evidence from US Trade Policy." *American Economic Review: Insights* 3 (1): 19–34.

Cavallo, Alberto, Paola Llamas, and Franco M. Vazquez. 2025. "Tracking the Short-Run Price Impact of U.S. Tariffs." NBER Working Paper 34496.

Chor, David, Matthew Grant, and Bingjing Li. 2025. "Exclusions for Sale? Tariff Exclusions in the US-China Trade War." Unpublished.

Cleveland, Robert B., William S. Cleveland, Jean E. McRae, and Irma Terpenning. 1990. "STL: A Seasonal-Trend Decomposition Procedure Based on Loess." *Journal of Official Statistics* 6 (1): 3–73.

Costinot, Arnaud, and Iván Werning. 2025. "How Tariffs Affect Trade Deficits." NBER Working Paper 33709.

Eck, Sydney, Trang Hoang, Carter Mix, and Madeleine Ray. 2026. "Mind the Gap: Announced versus Implied Tariff Rates in Recent Trade Policy Episodes." FEDS Notes, April 8. https://www.federalreserve.gov/econres/notes/feds-notes/mind-the-gap-announced-versus-implied-tariff-rates-in-recent-trade-policy-episodes-20260408.html.

Fajgelbaum, Pablo D., Pinelopi K. Goldberg, Patrick J. Kennedy, and Amit K. Khandelwal. 2020a. "The Return to Protectionism." *Quarterly Journal of Economics* 135 (1): 1–55.

Fajgelbaum, Pablo D., Pinelopi K. Goldberg, Patrick J. Kennedy, and Amit K. Khandelwal. 2020b. "Updates to Fajgelbaum et al. (2020) with 2019 Tariff Waves." Unpublished.

Fajgelbaum, Pablo D., and Amit Khandelwal. 2026. "Tariffs in 2025: Short-Run Impacts on the U.S. Economy." NBER Working Paper 35064.

Feenstra, Robert C., and Chang Hong. 2024. "Estimating the Regional Welfare Impact of Tariff Changes: Application to the United States." NBER Working Paper 33007.

Gopinath, Gita, Pierre-Olivier Gourinchas, Andrea F. Presbitero, and Petia Topalova. 2025. "Changing Global Linkages: A New Cold War?" *Journal of International Economics* 153: 104042.

Gopinath, Gita, Oleg Itskhoki, and Roberto Rigobon. 2010. "Currency Choice and Exchange Rate Pass-Through." *American Economic Review* 100 (1): 304–36.

Gopinath, Gita, and Brent Neiman. 2026. *Data and Code for: "The Incidence of US Tariffs: Rates and Reality."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E249900V1.

Hankins, Kristine W., Morteza Momeni, and David Sovich. 2026. "Consumer Credit and the Incidence of Tariffs: Evidence from the Auto Industry." *American Economic Review* 116 (2): 627–73.

Hayakawa, Kazunobu, Han-Sung Kim, and Taiyo Yoshimi. 2017. "Exchange Rate and Utilization of Free Trade Agreements: Focus on Rules of Origin." *Journal of International Money and Finance* 75: 93–108.

Hinz, Julian, Aaron Lohmann, Hendrik Mahlkow, and Anna Vorwig. 2026. "America's Own Goal: Who Pays the Tariffs?" Kiel Policy Brief 201.

International Monetary Fund. 2026. "Understanding Global Imbalances." IMF Policy Paper 2026/006.

Itskhoki, Oleg, and Dmitry Mukhin. 2025. "The Optimal Macro Tariff." NBER Working Paper 33839.

Johnson, Robert C., and Guillermo Noguera. 2012. "Accounting for Intermediates: Production Sharing and Trade in Value Added." *Journal of International Economics* 86 (2): 224–36.

Kimball, Miles S. 1995. "The Quantitative Analytics of the Basic Neomonetarist Model." *Journal of Money, Credit and Banking* 27 (4): 1241–77.

Klenow, Peter J., and Jonathan L. Willis. 2016. "Real Rigidities and Nominal Price Changes." *Economica* 83 (331): 443–72.

Minton, Robbie, and Mariano Somale. 2025. "Detecting Tariff Effects on Consumer Prices in Real Time." FEDS Notes, May 9. https://www.federalreserve.gov/econres/notes/feds-notes/detecting-tariff-effects-on-consumer-prices-in-real-time-20250509.html.

Shin, Hyun Song, Philip Wooldridge, and Dora Xia. 2025. "US Dollar's Slide in April 2025: The Role of FX Hedging." Bank for International Settlements Bulletin 105.

Werning, Iván, Guido Lorenzoni, and Veronica Guerrieri. 2025. "Tariffs as Cost-Push Shocks: Implications for Optimal Monetary Policy." NBER Working Paper 33772.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 145–170

Global Imbalances, Tariffs, and Industrial Policy

Pierre-Olivier Gourinchas, Gene Kindberg-Hanlon, Manasa Patnam, Lorenzo Rotunno, and Michele Ruta

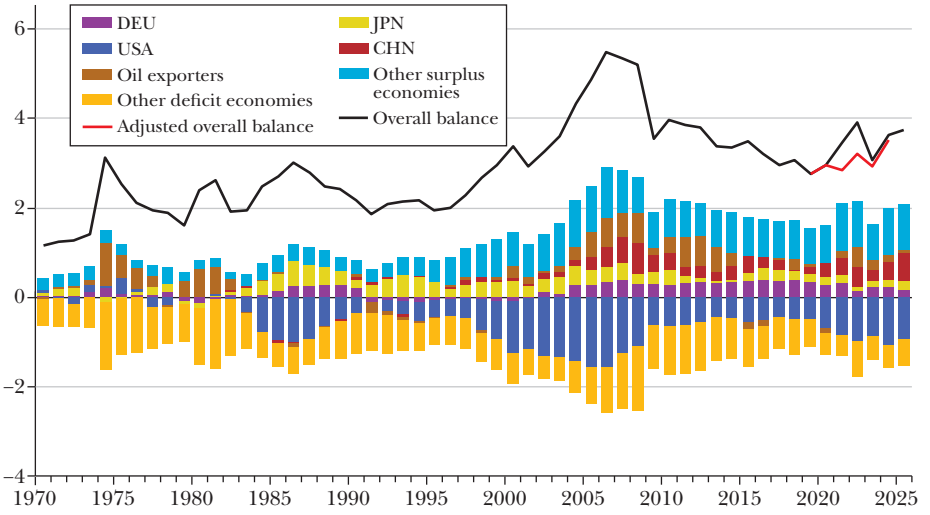
After a steep decline following the global financial crisis of 2008, global current account balances—measured as the sum of the absolute value of national current account surpluses and deficits—have widened since the pandemic (as illustrated in Figure 1), largely due to developments in the world's two largest economies, the United States and China (IMF 2026). The sources of this resurgence are the subject of an intense debate among economists and policymakers (Bai, Gopinath, Rey, and Weber 2026; Rey, Weder di Mauro, and Zettelmeyer 2026). This debate pits a traditional macroeconomic approach that discusses current account balances in the context of *intertemporal* saving and investment determinants against an alternative view focusing more narrowly on *intratemporal* drivers such as competitiveness and relative prices, arguing that imbalances can arise or be contained (as the case may be) with the help of industrial policies or import tariffs.

The standard macroeconomic approach notes that the current account balance of an economy is identically equal to the difference between two forward-looking

■ *Pierre-Olivier Gourinchas is Economic Counsellor and Director of Research at the IMF (International Monetary Fund, Washington DC), and S.K. and Angela Chan Professor of Global Management, Department of Economics and Haas School of Business, University of California, Berkeley (on leave). Gene Kindberg-Hanlon is Senior Economist at the IMF. Manasa Patnam is Deputy Division Chief at the IMF. Lorenzo Rotunno is Senior Economist at the IMF, and Professor of Economics at Aix-Marseille University (on leave). Michele Ruta is Division Chief at the IMF. The views in this paper do not necessarily reflect those of the IMF, IMF management, or the IMF executive board. Their emails are pog@berkeley.edu, gkindberg-hanlon@imf.org, mpatnam@imf.org, lrotunno@imf.org, and mruta@imf.org.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20261507.

Figure 1
Global Current Account Balances (Percent of World GDP)



Source: IMF, World Economic Outlook; IMF, External Sector Report; IMF, Balance of Payments.
Note: The overall balance is defined as the sum of absolute values of current account surpluses and deficits. Adjusted overall balance removes from the 2020–2024 headline global current account balance the impact of (1) COVID-19 pandemic factors and (2) commodity price fluctuations.

macroeconomic aggregates: national saving and domestic investment. Hence, changes in current account balances require understanding why and how intertemporal saving and investment decisions change in response to policies or shocks. This intertemporal approach emphasizes macroeconomic drivers—such as fiscal deficits in the United States that reduce US national saving, or rapidly aging populations and a property sector boom and bust in China that increase China's saving relative to its investment—as some of the key determinants of current accounts in both countries and globally (Obstfeld 2024; Gourinchas, Pazarbasioglu, Srinivasan, and Valdes 2024). This approach also emphasizes that current account deficits and surpluses are not necessarily negative outcomes. They can be consistent with economic fundamentals and desirable policies. For example, it is desirable for young or rapidly growing economies to finance part of their economic development with foreign capital and current account deficits. That said, large current account surpluses and deficits can pose risks when they are not justified by fundamentals. Growing financial vulnerabilities associated with increasing external balances often end with abrupt capital flow reversals, asset price corrections, and economic downturns.

The alternative view notes that the current account is also identically equal to the sum of the trade, primary income, and secondary income balances. Because,

in practice, the trade balance is often the largest component, this view focuses on the potential role of relative prices and national competitiveness. Proponents of this view often argue that China's external surpluses result from industrial policy measures designed to keep the currency artificially low to stimulate exports and support economic growth, amid weak domestic demand. The resulting surge in China's trade surpluses must be absorbed by trading partners, contributing to the erosion of their manufacturing base—a "China shock" (Yellen 2024; European Commission 2023; Pettis 2024).

Partly in response to rising concerns about these imbalances, other nations such as the United States have increased import tariffs and implemented industrial policies of their own. Recent examples of US industrial policy include major elements of the Infrastructure Investment and Jobs Act (2021), the CHIPS Act (2022), and the Inflation Reduction Act (2022). Since January 2025, the United States has also increased its import tariffs substantially, well above the average tariffs of other major economies, as shown in Figure 2, panel A. In addition, as panel B of Figure 2 shows, the use of industrial policy measures around the world, as recorded by the New Industrial Policy Observatory (Evenett et al. 2025), has surged in the last five years. While the number of measures implemented in China has surpassed those of other economies, industrial policies have also expanded more broadly across regions.[1] Available estimates place the cost of some of these policies such as direct and indirect subsidies at 4 percent of GDP per year in China (Garcia-Macia, Kothari, and Tao 2025) and up to 1 percent of gross output in other advanced and emerging economies (OECD 2025).[2]

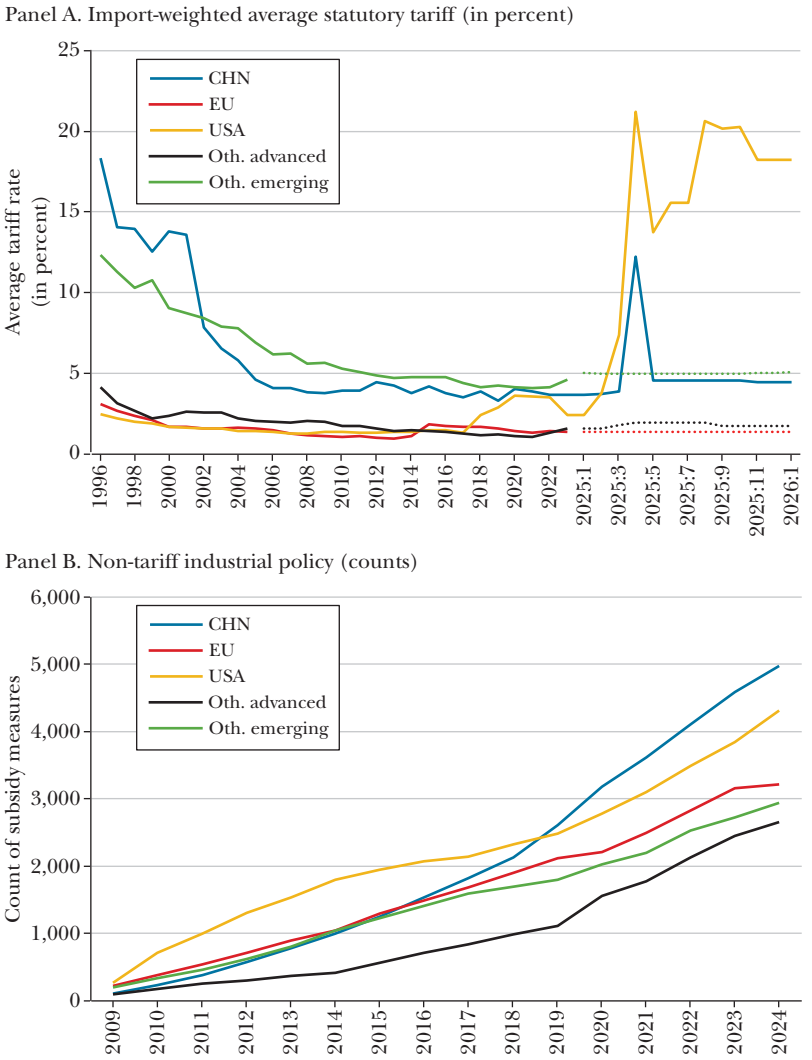
Obviously, the *inter-* and *intratemporal* approaches must ultimately agree. After all, national income accounting identities are always satisfied. Hence it is not enough to point to the potential competitiveness effects of tariffs, or to the increased incentive to produce tradable goods due to export-oriented industrial policies. These are partial equilibrium effects. What is needed is a macroeconomic general-equilibrium framework to assess how and through what channels these policy instruments may (or not) affect macroeconomic variables.

This paper provides a framework to help think about the effects of a range of policy tools such as macroeconomic and structural policies on current accounts. Conceptually, the determinants of the current account can be illustrated with the well-known stylized but pedagogical saving-investment Metzler diagram (Metzler

[1] Using different data sources and measurement approach, Ju, Li, and Wei (2026) confirm that industrial policies in the United States have increased sharply since 2017, but also find that this number may be up to four times higher than reported in the data used in Figure 2.

[2] Garcia-Macia, Kothari, and Tao (2025) assemble data from financial reporting of listed firms in China, complemented with information from other sources such as land registries. They include government grants, corporate income tax concessions, preferential access to credit (below-market borrowing), and subsidized land. Their estimated value of government support in 2023 is 3.9 percent of China's GDP (excluding subsidized land). The OECD MAGIC database (OECD 2025) collects data on industrial subsidies from the largest firms in key manufacturing sectors. They cover the same categories of government support as in Garcia-Macia, Kothari, and Tao (2025), except subsidized land. They report the value of subsidies as a share of firm revenues (see their Figure 1, panel B).

Figure 2
Increasing Use of Trade and Industrial Policy



Source: Panel A: Feodora Teti's Global Tariff Database (v_beta1-2024-12) from Teti (2024); BACI CEPII; ITC MACMAP; WTO-IMF Tariff Tracker; authors' calculations. Panel B: New Industrial Policy Observatory; authors' calculations.
Note: "Oth. Emerging" includes both emerging economies (excluding China) and low-income countries. "Oth. Advanced" includes advanced economies except the United States and the EU member states. Panel A: Import-weighted average for the European Union includes only external EU imports. Tariff data after 2024 are available only for the countries in the WTO-IMF tariff tracker as of January 2026 (Cambodia, Canada, China, India, Mexico, United Kingdom, United States, and Zimbabwe). Tariffs for other countries in 2025 (affecting the series and period in dashed lines) are kept at their 2024 level. Imports in 2024 are used as weights for tariffs in 2025 and 2026. Panel B: Cumulative sum of trade distortive industrial policy measures in force in a given year, excluding those classified as import policies. It includes only policies introduced since 2009, and hence misses measures introduced before that are still active during the sample period.

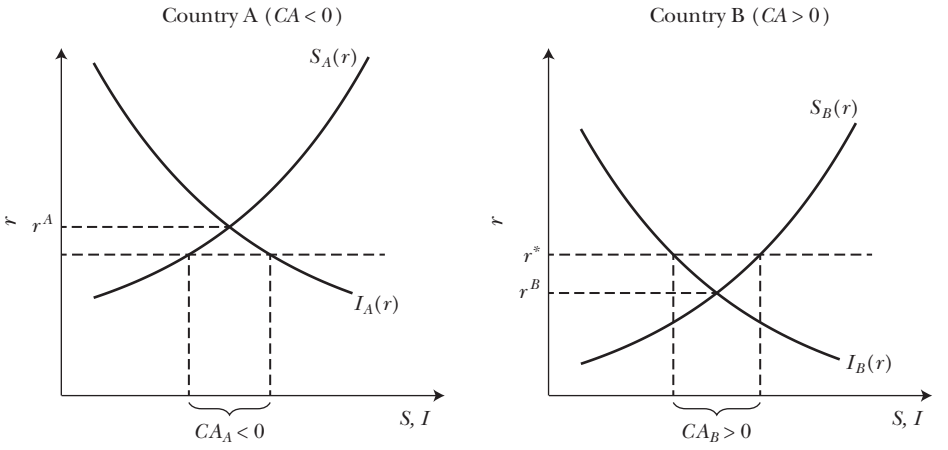
1960). We then extend this framework to consider the possible role of tariffs and industrial policy.

We argue that tariffs are a weak tool to respond to rising current account deficits. The literature has contrasted the effect of *temporary* tariffs, which can initially stimulate domestic saving by temporarily raising the price of imported goods, with *permanent* tariffs, which would leave intertemporal choices about saving and investment largely unchanged, in which case the current account must also be invariant to the tariffs. Instead, the standard theory concludes that the currency of the tariffing country should appreciate, depressing its exports and contributing to an unchanged trade balance. This is one of the clearest cases where the intertemporal approach delivers a superior insight: the appreciation of the tariffing country's currency is required precisely because intertemporal choices are unaffected. In more general settings, tariffs may reduce external deficits when tariff revenues are used to pay down public debt (increasing government saving), or when the initial external deficit is temporary (offset by future external surpluses). When instead external deficits are seen as permanent—reflecting large initial net foreign asset positions, or a convenience yield on domestic assets held abroad—the effects of tariffs on the current account are more complex and dictated by the structure of gross financial positions and of relative valuation effects, rather than changes in the relative price of imports and trade flows (Itskhoki and Mukhin 2025).

We also find that sector-specific industrial policies alone rarely boost current account balances, while broader forms of industrial policy can have more tangible impacts, but often at the expense of domestic consumption. Sectoral ("micro") industrial policies shift productive resources across industries. Because many such policies tend to be narrowly targeted and small in scope, they are unlikely to have significant effects at the aggregate level. An array of such policies could matter in the aggregate. Yet even then, their impact on the current account depends on whether they generate aggregate productivity improvements, temporary or permanent. Paradoxically, such measures tend to raise external surpluses only when they "fail" to durably boost competitiveness and aggregate productivity; for example, by creating persistent misallocation that lowers expected future income, reduces investment, and raises precautionary saving. Where these policies "succeed" in raising aggregate productivity, one should expect instead a decline in current account surpluses, as consumption increases in anticipation of higher future income and investments increase in anticipation of higher future returns to capital. By contrast, we argue that broader ("macro") industrial policies that include interventions such as capital controls, financial repression, or foreign exchange interventions—often implemented in combination with micro industrial policies—can significantly increase current account balances. They do so, however, by durably suppressing domestic consumption at home, hardly a recipe for sustained welfare improvements.

The exposition in this paper is intentionally stylized. However, in a few instances, we provide supporting model-based simulations of the effect of tariffs and industrial policies in various settings (see the accompanying Supplemental Appendix for more details). These simulations use a two-country New Keynesian model with

Figure 3
Saving-Investment Diagram of the Current Account



Source: Authors' creation.
Note: Saving-investment diagrams for two large regions, A and B. In A, the autarky interest rate r^A is assumed higher than the world interest rate r^*, whereas in B the autarky rate r^B is lower than r^*. Under free trade, region A runs a current account deficit ($CA_A < 0$) and region B runs a current account surplus ($CA_B > 0$).

traded and nontraded goods. The model features a range of common real-world frictions often included in medium-scale models.[3] A previous extended version of the paper contains a wider range of model results under different tariff, micro industrial policy and macro industrial policy scenarios (Gourinchas et al. 2026a).

A Primer on the Intertemporal Saving-Investment Framework

The determinants of global imbalances can be illustrated conceptually through the canonical textbook saving-investment Metzler diagram, which can be derived from the intertemporal theory of the current account (for the canonical model, see Obstfeld and Rogoff 1995). As illustrated in Figure 3, consider a world comprised of two regions and integrated capital markets, where saving is an increasing function of the real interest rate, while investments decrease with the real interest rate. The two regions are assumed to be "large" and hence their actions can affect the world interest rate. The current account balance for each country equals the difference between national saving and domestic investment.

[3]These include many that feature in other medium-scale models (such as Smets and Wouters 2007), including nominal rigidities, adjustment costs on investment, as well as hand-to-mouth credit constrained households and overlapping generations households—important features that increase the applicability of the model to real world macroeconomic dynamics.

In region A, residents are less frugal but have many investment opportunities, setting up a higher real interest rate in autarky, r^A. In region B, residents are inclined to save more and with a correspondingly lower real interest rate in autarky, r^B. When the two regions are financially integrated, the equilibrium global interest rate (r^*) lies between the two autarky interest rates and equates region A's desired borrowing (implying a current account deficit in region A) with region B's desired lending (implying a current account surplus in region B).

The diagram has been widely used in the research literature, and suffuses many policy discussions on the drivers of imbalances (as one prominent example, Bernanke's (2005) speech on the saving glut). A key determinant of how policies and macroeconomic shocks affect the current account balance is how they alter forward-looking saving and investment decisions. The intertemporal theory of the current account offers a useful framework to study the drivers of imbalances and how they can be rectified. For instance, a housing crash that increases uncertainty about future growth in region B can increase precautionary saving incentives and shift the saving curve further outward and desired investment inward. This would reduce world interest rates, and widen the current account surplus in region B and deficit in region A.

In the intertemporal framework, the persistence of a shock or policy matters for its effect on the current account, as does the directionality of its effect on aggregate variables. For example, a shock that temporarily raises the price of imports would have very different implications for short-term saving and investment decisions than one that permanently raised them. A subsidy policy that increases activity in industries with increasing returns to scale and raises future productivity would have similarly different effects than one increasing misallocation via costly tax rises or lower public service provision. In general, we expect the effects of policies and shocks on the current account to be larger when they vary over time. A policy whose effects are perceived to be time-invariant would be more neutral for the current account, because it would leave patterns of intertemporal substitution in saving and investment mostly unchanged.

Economists have used the saving-investment framework to assess the determinants of the current account balances both before and since the global financial crisis of 2008–2009 (for example, Bernanke 2005; Caballero, Farhi, and Gourinchas 2008; Blanchard and Milesi-Ferretti 2012). In the most recent episode of widening current account balances since the COVID pandemic, saving and investment drivers are likely to have been important factors for the world's two largest economies. In China, a large real estate downturn began at the end of 2021. This led to a sharp fall in real estate investment and a decline in consumer confidence. China's household saving rate increased sharply after the decline, potentially reflecting concern about future growth prospects and falling asset values. The household saving rate remains about 2 percentage points above its pre-COVID level. In the United States, the fiscal deficit has grown substantially relative to the pre-COVID period, even after COVID-related spending lapsed. As a result, the general government deficit is now about 2 percent of GDP higher than in 2017. The household saving rate has also

declined substantially since 2021, and consumption has grown more rapidly than expected, partly reflecting an unwind of "excess saving" accumulated during the pandemic as consumption was restricted. More recently, high equity valuations, potentially reflecting enthusiasm about future productivity growth in the tech and artificial intelligence sectors, are also likely to have contributed to lower household saving.

However, previous analysis has paid little attention to industrial and trade policies given their lower relevance to policy discussion at the time.[4] The intertemporal framework and quantitative models can be used to study the effects of tariffs and industrial policy on the current account, a task to which we now turn.

Tariffs and the Current Account

Throughout recent history, countries running current account deficits have at times resorted to import protection through tariffs with the objective of reducing imbalances. In recent years, the increased use of these policies has been framed in part as a response to perceived industrial policies and non-tariff barriers by trading partners that are considered responsible for worsening trade balances. Other proposed motivations for introducing tariffs include reshoring manufacturing employment, generating fiscal revenues, and eliminating bilateral trade balances. Our focus here is on the question of whether tariffs work to reduce aggregate imbalances. The effectiveness of import tariffs in reducing current account deficits has been the focus of recent theoretical and empirical papers (for example, Lorenzoni 2019; Schmitt-Grohé and Uribe 2025; Itskhoki and Mukhin 2025).

Intuitively, it may seem obvious that using tariffs to raise the price of imported products should discourage imports and lead to an improvement of the trade and current account balance, everything else equal. However, everything else is not equal. The intertemporal framework suggests that the effects of across-the-board tariffs on the current account are more nuanced, with the rise in import prices increasing national saving relative to investment only under certain circumstances.

Temporary Versus Permanent Tariffs and Deficits

To gain some intuition about the effects of tariffs on the current account balance, we consider the imposition of uniform tariffs by region A shown earlier in Figure 3—the deficit region—with the objective of reducing its initial current account deficit. The early literature on the effect of tariffs on the current account

[4] A recent paper by Itskhoki and Mukhin (2026) provides a similar overview of the drivers of global imbalances, with a focus on recent tariff actions. Conceptually, our saving-investment framework of the current account complements theirs—unsurprisingly then, we reach similar findings. Cesa-Bianchi, Ferrero, Fornaro, and Wolf (2026) instead assess the role of industrial policy and show that growth-enhancing government intervention can expand external surpluses when desired saving is constrained to rise with income and domestic assets are in short supply, extending the analysis in Caballero, Farhi, and Gourinchas (2008).

points out that the effect depends on the persistence of the tariffs shock. To appreciate the importance of this element, we can consider models without investment. In the case of temporary tariffs, domestic households face a higher real interest rate because consumption today—including imports—is more expensive than tomorrow. As a result, they are willing to increase saving and postpone consumption, improving the current account. In contrast, for a country without an initial trade deficit, a permanent tariff generates no such intertemporal trade-off. As tariffs will remain indefinitely higher, there is no incentive to adjust aggregate saving, thus leaving the current account unchanged.[5] Given an unchanged current account, the standard analysis concludes that the currency of the tariffing country must appreciate enough to offset the impact of the tariffs on relative prices and the trade balance (as shown in Razin and Svensson 1983), a finding known as Lerner (1936) symmetry. Importantly, this does not rule out real effects on imports and exports from the tariff. For example, imports will fall in region A as the tariff increases their cost. However, this creates a shortage of foreign exchange earnings in region B with which to purchase exports from region A. Because net saving is unchanged in region A, this shortage cannot be resolved through borrowing. Instead, excess demand for the currency of region A leads to an appreciation of the currency, a corresponding fall in exports, and a neutral current account effect.[6]

But what should happen if the country, like region A in our analysis, starts with an initial trade deficit? In that case, the analysis is more complex. We need to think about what happens when these trade deficits are transitory and expected to be offset by future trade surpluses so that the net present value of trade balances is zero, or if they are persistent—for example because they will be offset by net holdings of international assets and their associated income stream (Gourinchas and Rey 2007; Itskhoki and Mukhin 2026).

In the case of temporary trade deficits, permanent tariffs will reduce current account balances. The reason is both important and subtle. Because the country is running trade deficits today but will have trade surpluses at some future date, the composition of consumption changes over time and so does the consumption deflator, with more weights on (tariffed) imported goods when running trade deficits and less weight when running trade surpluses in the future. As a result, households in region A face a higher consumption-based real interest rate which discourages consumption and helps contain the current account deficit. Region B

[5] Recent work by Costinot and Werning (2025) qualifies this result, showing that it holds in a rather specific setting—for example, where tariffs do not lead to adjustments on the extensive margin of trade, or without terms-of-trade effects. In more general settings, permanent tariffs could narrow the deficits by increasing the marginal cost of consumption disproportionally more in the current period than in the future.

[6] Note that this mechanism applies under a system of floating exchange rates. In the case of fixed nominal exchange rates, there would be a slower-moving real exchange rate appreciation via domestic prices, which would result in an intertemporal tradeoff that increases saving and the current account in the near term (Erceg, Prestipino, and Raffo 2023).

faces a symmetric adjustment, with a lower consumption-based real interest rate, hence a smaller desire to save and a smaller current account.[7]

Now consider the situation in which countries can run permanent trade deficits. For example, this can happen in the cases where the deficit country starts with a strong net external creditor position, or where it earns higher returns on its investments abroad than what it pays on its liabilities abroad (as in the case of the "exorbitant privilege" of the US economy described in Gourinchas and Rey 2007), ensuring that a modest trade deficit is sustainable in the long run. In response to the imposition of permanent tariffs, the question becomes: what happens to the market value of cross-border assets, and to the net present value of excess returns? If these values are unchanged—admittedly an implausible assumption—we are back to the Lerner symmetry world and currencies must adjust so as to keep the present value of trade deficits unchanged. Otherwise, the change in the market value of the net foreign asset position and associated excess return revenues will dictate the change in the net present value of trade deficits, as emphasized by Itskhoki and Mukhin (2026).

The main insight here is that tariffs may have smaller or larger effects on saving and current account deficits depending on the initial deficit, net investment position, and equilibrium excess returns. However, the nature of the impact is uncertain and depends on perceptions of the durability of the deficit and on valuation effects on the investment position—and not on changes in relative import prices and trade patterns. Overall, this means that the partial equilibrium logic is a poor guide and tariffs will not reliably increase saving and the current account balance.

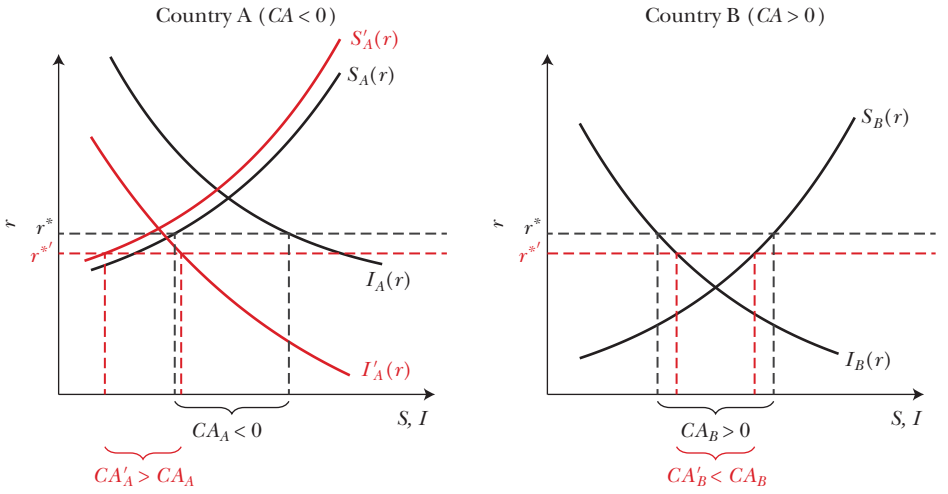
Investment Responses and Fiscal Implications

In models with investment, the effect of tariffs on the current account becomes yet more complex. Higher costs of imported goods could lower the marginal return to capital either because intermediate goods are imported, or because—through the tariff leading to a currency appreciation—the demand for exported goods falls. Such effects can overturn the basic "neutrality" result of the simpler intertemporal models (Sen and Turnovsky 1989).[8] Without a corresponding fall in desired saving, the current account will initially improve as investment declines. Furthermore, unilateral tariffs could also result in a fall in desired saving as a result of an improvement in the terms of trade or if tariff revenues are redistributed to households—in these cases, the real purchasing power for a given level of domestic output is higher, generating a lower saving rate relative to GDP. This situation is depicted in Figure 4—both the investment and saving schedules shift inward. In the

[7] See Dornbusch (1983) for a seminal analysis and Obstfeld and Rogoff (2001) for an early discussion of the effect of trade costs on current account balances. This result helps understand why, as tariffs keep increasing, trade flows collapse and in the limit countries revert to autarky as intuition would suggest.

[8] A special case where investment could instead increase occurs when tariffs induce a reallocation towards domestic investment and import competing industries, leading to a persistent increase in the current account deficit (Roldos 1991).

Figure 4
Effect of Tariffs on Imbalances



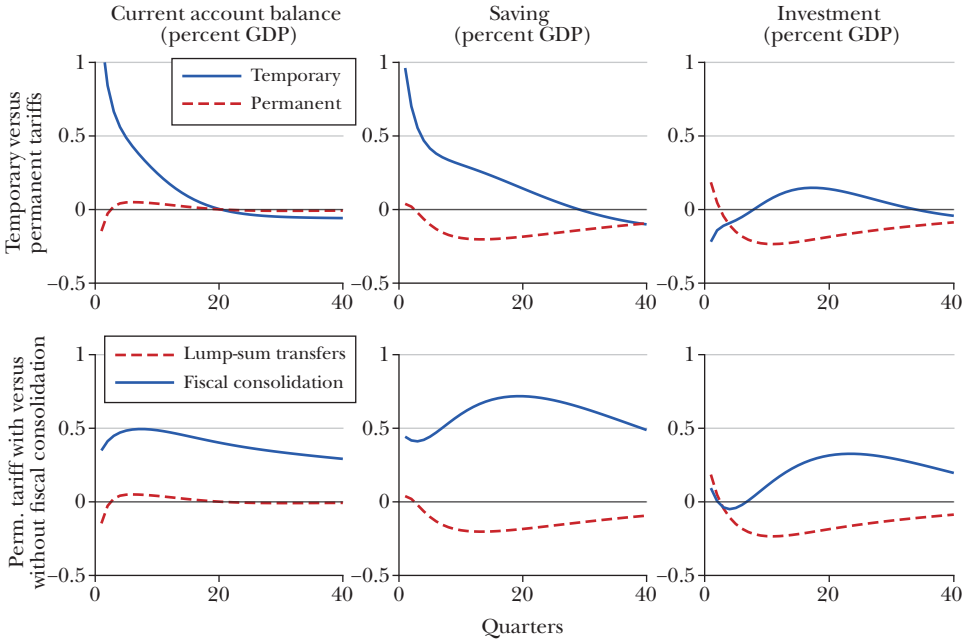
Source: Authors' creation.
Note: Saving-investment diagrams for regions *A* and *B*. In the counterfactual equilibrium, region *A* implements across-the-board import tariff increases perceived as permanent. The figure depicts a case where the tariff depresses desired investments as the cost of capital and imported inputs increases, and the saving schedule shifts inward—for example, because tariff revenues are redistributed to households. Because the shift in investment is assumed stronger than the one in saving, the deficit in *A* decreases, while the lower world interest rate also pushes down the surplus in region *B*.

specific example, the fall in investment dominates the reduction in saving, leading to a modest decrease in the current account deficit of country A.

Given the uncertainty surrounding the relative shifts of investment and saving in response to tariffs, we validate the intuitions gained from the Metzler framework using simulations from our two-country model mentioned in the introduction, with firms and households both reacting to current and expected output and price effects, while also subject to real world frictions. Overall, the results are consistent with the conceptual framework (Figure 5, top row). A temporary unilateral tariff generates a large increase in saving that offsets the decline in investment, generating a current account surplus. In the case of a permanent tariff, the decline in investment is closely matched by a fall in saving, as also shown in the Metzler diagram. This leaves the current account balance little changed overall.

The effects of tariffs on the current account also depends on their fiscal impact. When tariff revenues are not redistributed through some mixture of spending increases and tax cuts, but instead used to pay down government debt, the implications of tariffs for national saving and the current account are different. As long as households are non-Ricardian—that is, households do not react to reduced government borrowing by increasing their own borrowing by an offsetting amount—lower budget deficits increase overall national saving. In this case, the fiscal

Figure 5
Impacts of Temporary and Permanent Tariffs Under Different Treatments of Tariff Revenues



Source: Gourinchas et al. (2026b).
Note: Simulated as a unilateral 10 percent increase in tariffs on imports with all tariff revenues immediately distributed to households in the top row. In the temporary case, the tariffs have a half life of ten quarters. In the second row, tariff revenues are either fully immediately distributed to households, or only a small fraction is paid out each period so that most of the revenues pay down government debt (fiscal consolidation case).

consequence of the tariff generates a persistent increase in the current account, a case of twin surpluses (Figure 5, bottom row). In the case of a fiscal consolidation, interest rates are slightly lower, leaving investment slightly higher, but the improvement in the current account comes at the cost of a fall in output.

Empirically, the literature tends to find economically small effects of tariff shocks on the trade and current account balance, using both country-specific (US) data (Boer and Rieth 2024) and cross-country data (Furceri, Hannan, Ostry, and Rose 2022). Using historical US data, Schmitt-Grohé and Uribe (2025) find that, consistent with the predictions of the intertemporal framework, transitory tariff shocks increase the current account, while permanent changes have a muted effect. Quantitative analyses based on long-run general equilibrium trade models find that the US current account deficit as well as welfare decrease under policy scenarios that are based on US tariff increases during 2025 (Caliendo, Kortum, and Parro 2025; Ignatenko, Lashkaripour, Macedoni, and Simonovska 2025).

In short, tariffs are a limited tool for reducing current account deficits. Permanent tariffs typically fail to alter intertemporal saving decisions significantly, in which case the external balance remains largely unchanged. An improvement in the current account is more likely to arise if tariffs are temporary, thereby incentivizing saving, or if the initial deficit is seen as temporary. Additionally, the fiscal treatment of tariff revenue is critical: tariffs can narrow external deficits if proceeds are used to consolidate government debt, but this effect dissipates if revenues are redistributed through government spending or tax cuts.

Industrial Policy and the Current Account

Traditionally, industrial policies have been thought of as a set of government interventions that target certain industries or firms, generally with the goal of promoting growth in the targeted sectors for economic or non-economic reasons (Evenett et al. 2025). From an economic perspective, industrial policies can be desirable if/when they correct market failures: for instance, by increasing output in an industry characterized by external returns to scale or learning-by-doing. Any such policy, implemented in a narrow sector, is unlikely to have much of an aggregate effect. However, an array of such policies implemented across a broad range of sectors, or targeting large sectors of activity, could influence the current account balance via aggregate saving and investment decisions. These policies could include industrial subsidies across a large range of industries, but also broader macroeconomic policies that aim to promote industrial development and competitiveness through the deployment of more aggregate instruments such as capital controls, financial repression, and foreign exchange interventions. We refer to the latter group as "macro" industrial policy to distinguish them from the former more traditionally defined "micro" industrial policies.

The role of either kind of industrial policy in influencing external balances has been underexplored so far. There is a lack of research on the implications of sectoral or micro industrial policies for the current account, although some papers have emphasized the effects of subsidies and other government support on trade flows (Rotunno and Ruta 2024b; Huang et al. 2025). With regard to macro industrial policies, a few papers have investigated the impact on the current account of reserve accumulation interacted with restrictions on capital flows (Choi and Taylor 2022; Gagnon 2012), while the role of financial repression remains relatively understudied, also because of the difficulties in identifying and measuring the relevant government interventions.

Micro Industrial Policies and the Current Account: Limited Reach

Targeted Support: Productivity Effects. A key attribute of micro industrial policies is that when they seek to promote the favored firm or industry, it may be at the direct or indirect expense of other firms and industries. Such policies may include

"infant-industry" trade protection, production or export subsidies, directed or subsidized credits, or access to lower energy and other input prices from state-owned enterprises (financed by budgetary subventions or by cross-subsidies from other industries and households). The targeted nature of such interventions means that their primary effect is to reallocate productive resources across sectors and hence shape comparative advantage patterns, without clear effects on overall current account balances. However, traditional micro industrial policy in certain cases could be significant enough to impact aggregate saving and investment patterns, thus leading to changes in external balances. This can happen when the targeted sector is sufficiently large or has significant spillovers, perhaps due to the central role of the targeted sector in input-output networks (Liu 2019).

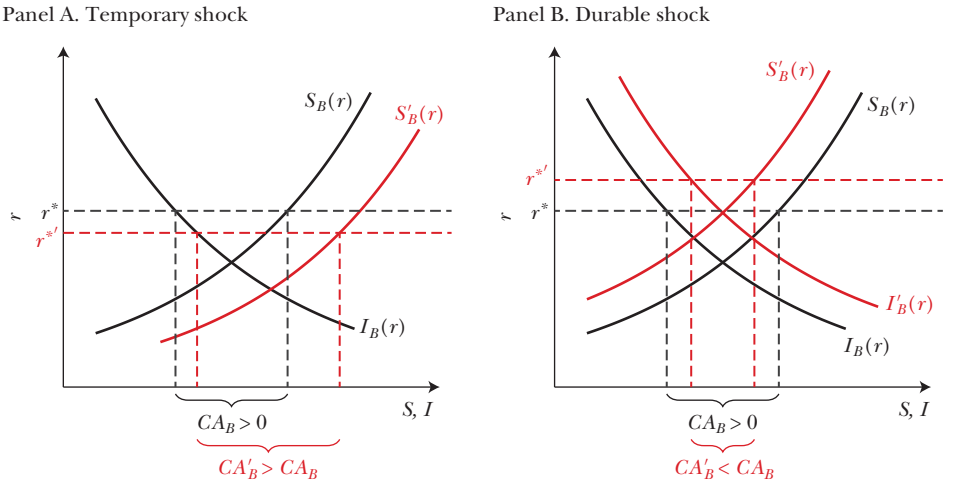
In this section, we focus on the outcome of micro industrial policy when sufficiently large to affect aggregate productivity. This effect could involve productivity gains if the industrial policy corrects significant market failures, or productivity losses through misallocation of resources. Certain industrial policy instruments such as restrictive licensing and regulations to favor certain industries may not have much fiscal impact, and so could affect current account balances only through this productivity channel.

While papers have shown how industrial policies can affect productivity for single countries and sectors—for example, in China (Garcia-Macia, Kothari, and Tao 2025); in the electric battery (Barwick, Kwon, Li, and Zahur 2025) and semiconductor (Goldberg et al. 2024) sectors—difficulties in measuring these policies consistently across countries and over time have prevented an empirical assessment of their effects on the current account. Reduced-form empirical studies suggest that subsidies and industrial policies expand exports and reduce imports in targeted products in China and other emerging economies (Rotunno and Ruta 2024*a*, b). However, quantitative analysis also suggests that these targeted policies can trigger important reallocation effects (Rotunno, Ruta, and Verma 2026), underscoring the ambiguity in any potential aggregate productivity effect.

To illustrate the possible effect of micro industrial policy through the channel of productivity effects, we assume that these measures are implemented by the surplus country (country B). To consider the productivity-only effects of industrial policy, we further assume that government intervention is operationalized through targeted fiscal instruments that are budget-neutral. Industrial policy can influence productivity to a macro-relevant degree either positively (for example, by shifting more resources to sectors characterized by increasing returns to scale and learning-by-doing externalities) or negatively (by shifting resources away from sectors with positive externalities and high levels of productivity).

The direction of any productivity effect on the current account further depends on whether the productivity effects are perceived as transitory or permanent, similar to the tariff case. The graphs in the two panels of Figure 6 depict these two cases in a scenario where industrial policy affects productivity positively. In panel A, a temporary industrial policy-led rise in productivity in country B (for example, when agents expect the positive effects to be short-lived) shifts the saving schedule to the

Figure 6
A Positive Micro Industrial Policy-Induced Productivity Shock and Imbalances



Source: Authors' creation.
Note: Saving-investment diagrams for region *B*. Red colored terms and lines denote a counterfactual equilibrium under higher (micro-industrial policy induced) productivity. In panel A, the productivity shock is assumed temporary. The saving schedule shifts outward as households smooth the increase in current income over time. The current account surplus in *B* increases ($CA'_B > CA_B$) and the world interest rate goes down. In panel B, expected productivity is assumed to increase durably. The investment schedule shifts out, while desired saving decreases because of higher future income. As a result, the current account surplus in *B* shrinks ($CA'_B < CA_B$) and the world interest rate goes up.

right as agents smooth the increase in current income over time. The real interest rate adjusts downwards, and country A increases borrowing—the current account surplus of country B widens.

In panel B, an industrial policy-induced increase in expected productivity would shift the investment schedule to the right and the saving schedule to the left because future income is higher than current income. As a result, country B's autarky rate and the world interest rate increase. Faced with a higher interest rate, country A's deficit and country B's surplus would decrease. Finally, consider the case where industrial policy results in a lower aggregate productivity, for example through an increased misallocation of resources. In this situation, the surplus of country B would initially increase if the productivity effects of industrial policies are perceived as durable. This analysis suggests that industrial policies are expected to increase current account surpluses only when they are unsuccessful—that is, if they increase aggregate productivity only temporarily, or if they lead to structural misallocation that decreases aggregate productivity durably.

In addition, the sector that industrial policy supports will matter. When industrial policy supports the tradable sector but simultaneously leads to a misallocation of resources that lowers productivity in the nontradable sector, the outcome is not

simply a mirror image of the positive productivity case (Glick and Rogoff 1995; Stockman and Tesar 1995). A durable decline in nontradable productivity can weaken the current account balance just as an improvement in tradables productivity can, but through a different channel. In this scenario, the fall in nontradable sector productivity leads to a scarcity of these products, prompting rising relative prices in the nontradable sector because they are poor substitutes for tradable products (see Gourinchas et al. (2026a) and the Supplemental Appendix for further details). The result is an appreciation of the real exchange rate and loss of competitiveness, further decreasing the current account balance.

Targeted Support: Boosting Output with Subsidies or Targets. We now consider the case of a micro industrial policy implemented through subsidies to investment that have a fiscal effect or through directives that set quantity targets with the aim of "bending the cost curve"—in this context, this goal refers to achieving scale economies that bring down marginal costs. Again, we can imagine these policies as partially targeted, but of sufficiently large scale to have aggregate implications on tradable sector investment and prices. However, here we do not consider second-round effects on productivity as in the previous section. We discuss the implications of these policies verbally and refer to the Supplemental Appendix for a detailed presentation of the model-based simulations.

Take the example of a permanent subsidy to investment in the tradable sector. As in the case of productivity-enhancing industrial policy, such a subsidy will increase investment and then output in the future through capital deepening. In anticipation of higher income, households initially dissave. Both the higher domestic investment and lower domestic saving will reduce the current account balance. However, the source of the funds for the subsidy now becomes important. Higher taxes to fund the subsidy reduce disposable income and therefore limit the degree of initial dissaving relative to the deficit-financed case. The situation would be as the one depicted in panel B of Figure 6, but with an attenuated inward shift in the saving schedule. This effect is especially strong when the fiscal burden falls on credit-constrained households, whose consumption responds sharply to current disposable income. The current account balance still declines, but by less when the subsidy is financed through immediate taxation rather than through higher deficits.

We next consider the case of policies to boost quantities in an export-focused (tradable) sector but without the use of subsidies. These policies could include state directives to achieve desired output targets, or other similar directives to increase output that may override commercial interests. These interventions have similar effects to the productivity improvement case considered above—an increase in production to the detriment of profit margins is a positive supply shock for the tradable sector. Investment rises rapidly as firms target higher output at lower margins. At the same time, a relative shortage of inputs in the nontradable sector creates additional demand for more workers and investment there, raising prices and wages overall. The result of this rise in domestic costs and income is a sharp appreciation of the real exchange rate, contributing to reverse any expansion in net

exports as a direct effect of higher output levels in the tradable sector. This mechanism resembles a Balassa-Samuelson-type channel (Balassa 1964), in that pressures in the tradable sector raise economy-wide costs and appreciate the real exchange rate, although here the driver is a policy-induced expansion in output rather than productivity gains. The net effect can easily be that, while the policy targets a larger output in the tradable sector, the current account balance falls through income and real exchange rate effects.

The overall conclusion is that durable and successful micro-style industrial policies are unlikely to increase the current account balance. If they succeed in boosting productivity or investment and output in the export-focused sector, they will drive domestic investment and consumption higher in anticipation of income gains while also raising overall prices as they compete for resources with the nontradable sector, leading to a real currency appreciation. Overall, the current account balance will decline. If they fail, and instead generate misallocation that harms the nontradable sector, they will increase domestic price pressures and wages and decrease competitiveness, weighing on the current account as well. The only case in which they will boost the current account balance is where they have a harmful effect on productivity in the sector they are designed to help (for example, the export sector), or only temporarily increase productivity.

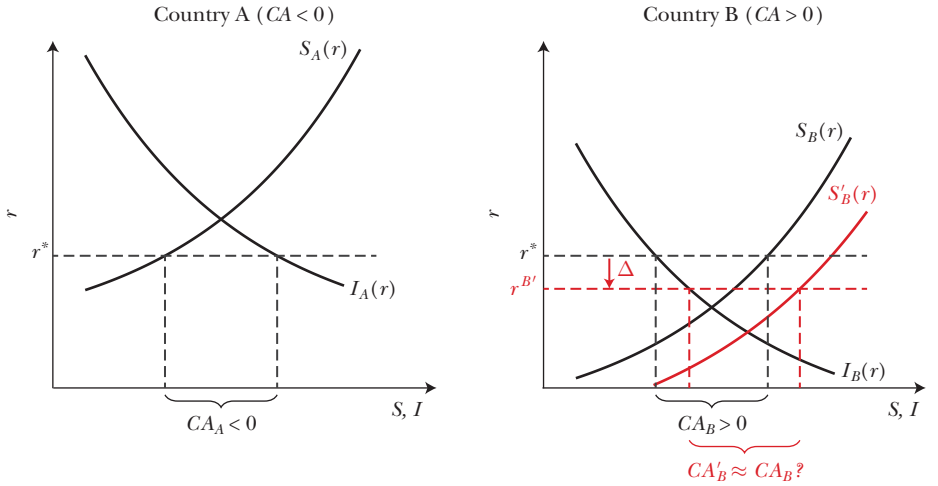
Macro Industrial Policies: Broader Reach, More Bite

Financial Repression and Forced Saving. While we often think of industrial policies as measures that are targeted to specific industries, certain macro and financial policies are sometimes deployed in conjunction with targeted industrial policies with the goal of promoting industrial development and exports on a broad scale. This section considers policies that seek a general decline in the cost of borrowing through financial repression or forced saving. The next section considers policies that seek to improve export competitiveness and promote import substitution by directly manipulating the capital account.

The term "financial repression" refers to a group of policies. It may involve limiting the interest rates available on deposits in the domestic banking sector or limiting the availability of other financial assets, in order to reduce funding costs, while simultaneously limiting lending rates on loans to businesses in targeted sectors. These policies are often combined with forced saving policies that can include prohibitions on dividend payments by corporations (to encourage higher retained earnings), reserve requirements on export earnings, mandatory saving requirements in pensions or similar products, or even weak social safety net provision that requires high private saving for insurance purposes.

Governments sometimes undertake financial repression measures for competitiveness reasons—that is, to reduce interest rates and increase credit allocation, with the goal of artificially lowering the cost of capital for the industrial or export sectors (for example, McKinnon 1973). In these cases, the intervention has an industrial policy intent. However, financial repression measures can also be

Figure 7
Effects of Financial Repression and Capital Flow Measures



Source: Authors' creation.
Note: Saving-investment diagrams for regions *A* and *B*. In the counterfactual equilibrium, region B implements a combination of forced saving (shifting the saving schedule of *B* outward) and restrictions to capital outflows (lowering the local interest rate $r^{B'}$ below the world interest rate r^*). The capital controls create a wedge Δr between the interest rates in region *A* and region *B*. The figure depicts a case where the current account balances remain unchanged as the interest rate wedge in Country B sufficiently increases investment and lowers saving enough to offset the schedule shift.

used for budgetary purposes—that is, with the objective of controlling or reducing interest payments on government debt.

As in the case of sectoral industrial policies, empirical evidence on the effects of financial repression measures on the current account balance is scant. Measurement issues are pervasive, given the difficulty in applying a common definition of financial repression policies across countries and over time. Abiad, Detragiache, and Tressel (2010) construct an index encompassing seven financial repression policies until 2005—including interest rate controls, barriers to entry in the financial sectors, and capital account restrictions. The index was extended to 2017 by Jafarov, Maino, and Pani (2019), but only for the component related to interest rate controls. Using this index, Johansson and Wang (2012) find a positive association between the current account balance and financial repression across countries.

Because the goal of these policies is to lower the domestic cost of capital relative to global interest rates, these measures are typically associated with restrictions on capital outflows. In the context of Figure 7, this creates a wedge $\Delta = r^{A'} - r^{B'} > 0$ between the interest rates in region B and abroad, as agents in region B face barriers to purchase or sell foreign assets. The result is that forced saving measures allow a larger decline in domestic interest rates and a corresponding boost to domestic investment.

The simultaneous increase in domestic saving and domestic investment in region B has an ambiguous effect on the current account of both regions. Figure 7 depicts the knife-edge case where the combination of forced saving and capital controls in B leaves the world interest rate (which prevails in region A) unchanged. As a result, the current account balances remain unaffected in both regions. In a case without capital controls, forced saving policies in country B would unambiguously increase its current account balance as global capital markets provide a release valve for higher domestic saving. The result would be a more modest increase in domestic investment. Analytically, this hypothetical case is close to the "saving glut" argument put forward in the early 2000s, when different structural and policy factors led to a global increase in the desired supply of saving, widening external balances and driving down interest rates (Bernanke 2005; Caballero, Farhi, and Gourinchas 2008).

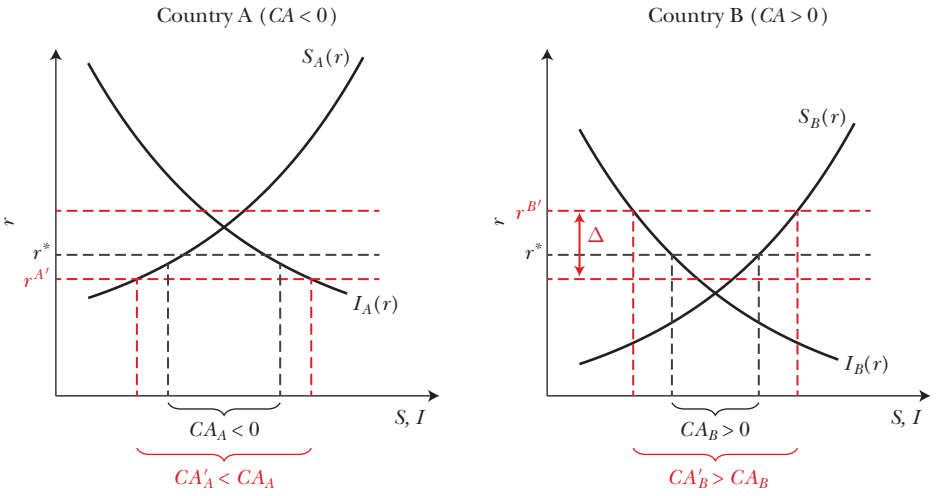
Capital Account Policies and Reserve Accumulation. Another "macro" industrial policy consists in using foreign exchange policy to manipulate the capital account. By the balance of payments identity, the current account balance must equal a corresponding inflow or outflow of capital—known as the "capital account." Hence a policy to manipulate the capital account must generate a current account of the same size. For the rest of the world to absorb the corresponding current account surplus, the real exchange rate must depreciate. One way to achieve this goal is to accumulate foreign reserves (Korinek and Servén 2016).[9] Some imperfect international capital mobility (perhaps created by government policy) is needed to ensure that domestic households cannot perfectly offset government reserve accumulation by selling foreign assets.

In the limit case where the capital account is closed for private investors, the path of reserve accumulation by region B exactly determines the current account balance, as shown by Jeanne (2013). However, in order to finance the accumulation of foreign exchange reserves, authorities must either raise taxes or issue domestic bonds. Either way, national saving must rise. Controls on capital inflows can increase the domestic interest rate, while the interest in region A falls because of the additional capital inflows in the form of reserves. Therefore, a wedge emerges between the real interest rates in both regions, but in contrast to the financial repression case, this time interest rates rise relative to rates abroad, $\Delta = r^{A'} - r^{B'} < 0$. While the current account surplus increases in B, higher interest rates depress investment. As shown in the Metzler diagram of Figure 8, the lower interest rate in A pushes down saving and hence widens the current account deficit.

Empirical work confirms the prediction that reserve accumulation coupled with restriction to capital inflows can increase the current account. Cross-country regressions using different instruments to isolate exogenous drivers of changes in

[9] This "mercantilist" view of foreign reserve accumulation is different from a "precautionary" motive by which reserves are accumulated to provide insurance against unexpected output losses in the case of a "sudden stop" crisis (Jeanne and Rancière 2011).

Figure 8
Effects of Foreign Reserve Accumulation and Capital Controls



Source: Authors' creation.
Note: Saving-investment diagrams for regions *A* and *B*. In the counterfactual equilibrium, region *B* implements a combination of reserve accumulation and restrictions to capital inflows. These policies create a wedge between the interest rates of the two countries—in region *B*, the interest rate $r^{B'}$ is higher than the world interest rate, while in region *A* the interest rate decreases. The current account surplus increases in *B*, while the deficit widens in *A*.

official reserves find that the current account balance is higher in countries with larger increases in reserves and more restrictive capital controls (Bayoumi, Gagnon, and Saborowski 2015; Phillips et al. 2013; Choi and Taylor 2022).

In sum, does the argument that macro industrial policies such as financial repression and exchange rate interventions worsen trade imbalances hold water? The short answer suggested by the theory and evidence is "only in specific circumstances." What the government can achieve is higher saving through financial repression policies, which in turn will increase the saving-investment balance. This policy would generate similar outcomes to the "saving glut" of the 2000s, lowering global interest rates and increasing imbalances. However, if this policy is combined with capital control policies to "bottle-up" saving domestically in order to lower domestic interest rates below global rates, it will limit the accumulation of foreign assets and thus lead to an improvement in the current account. The government can pursue policies of steady accumulation of foreign exchange reserves, but in contrast to the financial repression case, this requires capital controls to enable authorities to control the capital account and thus the current account balance. Both sets of policies also cause adverse economic outcomes: the suppression of consumption in the case of financial repression/forced saving, and higher domestic interest rates and lower investment in the case of reserve accumulation.

Combining Macro and Micro Industrial Policy

What about a combination of micro and macro industrial policies? As we have argued, "micro" industrial policies that successfully and persistently increase tradable sector output do not result in an increased current account balance. The income effects associated with successful policies raise consumption and imports, while the fundamental imbalance they create in the economy with the nontradable sector partially reverses the competitiveness improvements driven by the policy, appreciating the real exchange rate and overheating the economy (Obstfeld 2026). However, in combination with the macroeconomic industrial policies such as forced saving, these side effects could be contained, resulting in an improvement in the current account balance, output, and competitiveness. The cost of this combination of policies is that internal balance is restored through suppressed consumption.

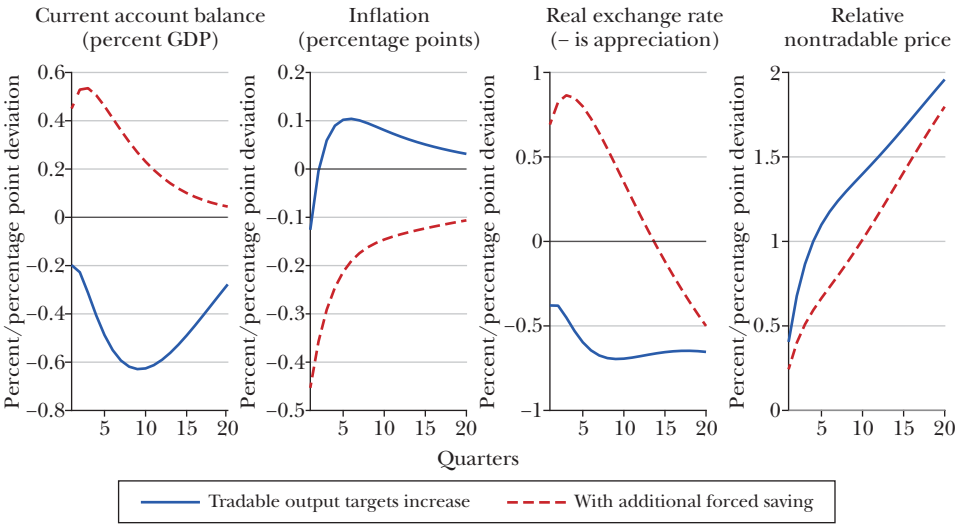
Consider the example of policy directives to boost tradable output discussed earlier, perhaps through the use of quantity targets. The policy increases tradable output, but also creates a lopsided economy, resulting in a shortage of nontradable goods. To restore balance, the price of domestic nontradable goods would rise rapidly, increasing the overall price level and appreciating the real exchange rate, as the model-based simulations in Figure 9 show. However, these side-effects of the policy can be reversed by suppressing consumption, for instance through forced saving policies. In this case, reduced aggregate demand contains the rapid rise in nontradable prices and overall inflation.[10] A sufficient degree of consumption repression can fully reverse the fall in saving and appreciation of the real exchange rate that would otherwise occur, subdue the relative nontradable price increase, and turn the current account balance effect from negative to positive.[11] The same principle can be applied to other forms of micro industrial policy, where productivity gains and subsidy policies can be combined with forced saving to both increase output and the current account balance.

It is worth emphasizing that the real exchange rate movements in these scenarios are equilibrium outcomes tied to underlying saving-investment imbalances. In particular, policies that raise aggregate saving, such as forced saving, generate current account surpluses which require a real depreciation in equilibrium. By contrast, a purely nominal exchange rate adjustment, absent changes in saving behavior, would not deliver a sustained *real* depreciation. Higher import prices would feed into domestic inflation, offsetting the initial competitiveness gains. Similarly, a policy that would seek to appreciate the nominal exchange rate of surplus countries

[10] Cesa-Bianchi, Ferrero, Fornaro, and Wolf (2026) come to a similar conclusion, although in their framework, consumption suppression arises endogenously alongside support for the tradable sector. Domestic bonds are assumed to provide utility but are in limited supply. As income rises following the implementation of industrial policy, households are assumed to increase demand for liquid assets (above-and-beyond demand implied for consumption smoothing), turning to foreign assets to meet demand and increasing the current account balance.

[11] This echoes an observation by Gerschenkron (1962) that industrialization in late-developing economies such as Russia and Germany systematically required holding back consumption to free up resources for capital-intensive goods sectors.

Figure 9
Effects of Combined Macro and Micro Industrial Policy on Prices and the Current Account



Source: Gourinchas et al. (2026b).
Note: "Tradable output targets increase" simulated as a permanent fall in the desired markup for tradable sector firms, which provides a mechanism to increase output in the sector at the cost of margins and profitability in the sector. In the "with additional forced saving" scenario, an additional shock to intertemporal preferences lowers consumption.

without a change in the underlying drivers of saving and investment would only result in stronger domestic deflationary pressures. In this sense, focusing on the nominal exchange rate alone risks overstating its ability to correct external imbalances without accompanying changes in domestic demand and saving.

Conclusion

This paper has developed a framework to think about the role of tariffs and industrial policies in driving current account balances. Importantly, it lays out the conditions where these policies work to improve balances (often at the expense of output and consumption), and importantly, when they will fail and even generate the opposite effect. We analyze the role of tariffs as a tool to reduce current account deficits, and of industrial policy as a determinant of current account surpluses. We consider different forms of industrial policy in our analysis: a more traditional sector-specific policy of support via subsidies or other targeted instruments (micro industrial policy), and broader policies of support through macroeconomic policies targeting interest rates with induced effects on the exchange rate (macro industrial policy).

Both tariffs and micro industrial policies have ambiguous effects on current account balances unless they are only temporary, when they generally increase current account balances. If tariffs and micro industrial policies are permanent, they can only increase the current account balance at the cost of output (for example, productivity-lowering misallocation), or the suppression of consumption (for example, using tariff revenues for fiscal consolidation). Macro industrial policies have a clearer impact on the current account, with forced saving and currency undervaluation unambiguously boosting the current account. However, these policies also suppress consumption to improve the current account balance.

Because it is difficult to observe the extent to which industrial policies have either underperformed or been combined with such macro-level demand-depressing measures, quantifying their contribution to widening current account balances is inherently challenging. With this caveat in mind, industrial policies are unlikely to be the only driver of the recent widening in global current account balances. Domestic factors related to the real estate downturn in China and widening fiscal deficits in the United States emerge as plausible drivers of at least some of the recent widening in balances (Gourinchas, Pazarbasioglu, Srinivasan, and Valdes 2024). A combination of macro and micro industrial policies may have played a role in the emergence of large external surpluses for China where a combination of large subsidy and financial policies has been implemented. But as our analysis shows, they do so at the cost of severely distorting the domestic economy. The message is clear: addressing domestic imbalances—weak domestic demand and misallocation of resources in China, excessive public deficits in the United States—remains the surest way to reduce imbalances. More industrial policies or tariffs might find these increases hard to correct.

■ *The authors would like to thank Joseph Gagnon, Olivier Jeanne, Gian Maria Milesi-Ferretti, Maurice Obstfeld, Zheng Song, Jeromin Zettelmeyer, and participants at the Bruegel at 20 meeting on Global Imbalances and at seminars hosted by the European Commission, the European Central Bank, the OECD, and the IMF.*

References

Abiad, Abdul, Enrica Detragiache, and Thierry Tressel. 2010. "A New Database of Financial Reforms." *IMF Staff Papers* 57 (2): 281–302.

Bai, Chong-En, Gita Gopinath, Hélène Rey, and Axel Weber. 2026. "G7 Economists Memo on Global Imbalances." Unpublished.

Balassa, Bela. 1964. "The Purchasing-Power Parity Doctrine: A Reappraisal." *Journal of Political Economy* 72 (6): 584–96.

Barwick, Panle Jia, Hyuk-Soo Kwon, Shanjun Li, and Nahim B. Zahur. 2025. "Drive Down the Cost: Learning by Doing and Government Policies in the Global EV Battery Industry." NBER Working Paper 33378.

Bayoumi, Tamim, Joseph Gagnon, and Christian Saborowski. 2015. "Official Financial Flows, Capital Mobility, and Global Imbalances." *Journal of International Money and Finance* 52: 146–74.

Bernanke, Ben S. 2005. "The Global Saving Glut and the US Current Account Deficit." Sandrige Lecture, Virginia Association of Economics, Richmond, VA, March 10. https://www.federalreserve.gov/boarddocs/speeches/2005/200503102/.

Blanchard, Olivier, and Gian Maria Milesi-Ferretti. 2012. "(Why) Should Current Account Balances Be Reduced?" *IMF Economic Review* 60 (1): 139–50.

Boer, Lukas, and Malte Rieth. 2024. "The Macroeconomic Consequences of Import Tariffs and Trade Policy Uncertainty." IMF Working Paper 24/13.

Caballero, Ricardo J., Emmanuel Farhi, and Pierre-Olivier Gourinchas. 2008. "An Equilibrium Model of 'Global Imbalances' and Low Interest Rates." *American Economic Review* 98 (1): 358–93.

Caliendo, Lorenzo, Samuel S. Kortum, and Fernando Parro. 2025. "Tariffs and Trade Deficits." NBER Working Paper 34003.

Cesa-Bianchi, Ambrogio, Andrea Ferrero, Luca Fornaro, and Martin Wolf. 2026. "Industrial Policies, Global Imbalances, and Technological Hegemony." Unpublished.

Choi, Woo Jin, and Alan M. Taylor. 2022. "Precaution versus Mercantilism: Reserve Accumulation, Capital Controls, and the Real Exchange Rate." *Journal of International Economics* 139: 103649.

Costinot, Arnaud and Iván Werning. 2025. "How Tariffs Affect Trade Deficits." NBER Working Paper 33709.

Dornbusch, Rudiger. 1983. "Real Interest Rates, Home Goods, and Optimal External Borrowing." *Journal of Political Economy* 91 (1): 141–53.

Erceg, Christopher, Andrea Prestipino, and Andrea Raffo. 2023. "Trade Policies and Fiscal Devaluations." *American Economic Journal: Macroeconomics* 15 (4): 104–40.

European Commission. 2023. "Statement by President von der Leyen at the Joint Press Conference with President Michel Following the EU-China Summit." Statement, December 6. https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_23_6409.

Evenett, Simon, Adam Jakubik, Jaden Kim, Fernando Martín, Samuel Pienknagura, Michele Ruta, Sandra Baquie, Yueling Huang, and Rafael Machado Parente. 2025. "Industrial Policy since the Great Financial Crisis." IMF Working Paper 25/222.

Furceri, Davide, Swarnali A. Hannan, Jonathan D. Ostry, and Andrew K. Rose. 2022. "The Macroeconomy after Tariffs." *World Bank Economic Review* 36 (2): 361–81.

Gagnon, Joseph. 2012. "Global Imbalances and Foreign Asset Expansion by Developing Economy Central Banks." Peterson Institute for International Economics Working Paper 12-5.

Garcia-Macia, Daniel, Siddharth Kothari, and Yifan Tao. 2025. "Industrial Policy in china: Quantification and Impact on Misallocation." IMF Working Paper 25/155.

Gerschenkron, Alexander. 1962. *Economic Backwardness in Historical Perspective.* Belknap Press, Harvard University Press.

Glick, Reuven, and Kenneth Rogoff. 1995. "Global versus Country-Specific Productivity Shocks and the Current Account." *Journal of Monetary Economics* 35 (1): 159–92.

Goldberg, Pinelopi K., Réka Juhász, Nathan J. Lane, Giulia Lo Forte, and Jeff Thurk. 2024. "Industrial Policy in the Global Semiconductor Sector." NBER Working Paper 32651.

Gourinchas, Pierre-Olivier, Gene Kindberg-Hanlon, Manasa Patnam, Lorenzo Rotunno, and Michele Ruta. 2026a. "Global Imbalances, Industrial Policy and Tariffs." IMF Working Paper 26/67.

Gourinchas, Pierre-Olivier, Gene Kindberg-Hanlon, Manasa Patnam, Lorenzo Rotunno, and Michele Ruta. 2026b. *Data and Code for: "Global Imbalances, Tariffs, and Industrial Policy."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E249860V1.

Gourinchas, Pierre-Olivier, Ceyla Pazarbasioglu, Krishna Srinivasan, and Rodrigo Valdes. 2024. "Trade Balances in China and the US Are Largely Driven by Domestic Macro Forces." IMF Blog, September 12. https://www.imf.org/en/blogs/articles/2024/09/12/trade-balances-in-china-and-the-us-are-largely-driven-by-domestic-macro-forces.

Gourinchas, Pierre-Olivier, and Hélène Rey. 2007. "International Financial Adjustment." *Journal of Political Economy* 115 (4): 665–703.

Huang, Yueling, Sandra Baquie, Florence Jaumotte, Jaden Kim, Yucheng Lu, Rafael Machado Parente, and Samuel Pienknagura. 2025. "Do Industrial Policies Increase Trade Competitiveness?" IMF Working Paper 25/98.

Ignatenko, Anna, Ahmad Lashkaripour, Luca Macedoni, and Ina Simonovska. 2025. "Making America Great Again? The Economic Impacts of Liberation Day Tariffs." *Journal of International Economics* 157: 104138.

IMF. 2026. "Understanding Global Imbalances." IMF Policy Paper 2026/006.

Itskhoki, Oleg, and Dmitry Mukhin. 2025. "The Optimal Macro Tariff." NBER Working Paper 33839.

Itskhoki, Oleg, and Dmitry Mukhin. 2026. "Global Imbalances: A Progress Report." In *Paris Report 4: The New Global Imbalances*, edited by Hélène Rey, Beatrice Weder di Mauro, and Jeromin Zettelmeyer, 145–69. CEPR Press.

Jafarov, Etibar, Rodolfo Maino, and Marco Pani. 2019. "Financial Repression Is Knocking at the Door, Again." IMF Working Paper 19/211.

Jeanne, Olivier. 2013. "Capital Account Policies and the Real Exchange Rate." In *NBER International Seminar on Macroeconomics*, Vol. 9, edited by Francesco Giavazzi and Kenneth West, 7–42. University of Chicago Press.

Jeanne, Olivier, and Romain Rancière. 2011. "The Optimal Level of International Reserves for Emerging Market Countries: A New Formula and Some Applications." *Economic Journal* 121 (555): 905–30.

Johansson, Anders C., and Xun Wang. 2012. "Financial Repression and External Imbalances." Stockholm School of Economics CERC Working Paper 20.

Ju, Jiandong, Yuankun Li, and Shang-Jin Wei. 2026. "The United States as an Active Industrial Policy Nation." NBER Working Paper 34744.

Korinek, Anton, and Luis Servén. 2016. "Undervaluation through Foreign Reserve Accumulation: Static Losses, Dynamic Gains." *Journal of International Money and Finance* 64: 104–36.

Lerner, Abba P. 1936. "The symmetry between Import and Export Taxes." *Economica* 3 (11): 306–13.

Liu, Ernest. 2019. "Industrial Policies in Production Networks." *Quarterly Journal of Economics* 134 (4): 1883–948.

Lorenzoni, Guido. 2019. "Do Tariffs Reduce the Trade Deficit?" Unpublished.

McKinnon, Ronald I. 1973. *Money and Capital in Economic Development*. Brookings Institution Press.

Metzler, Lloyd A. 1960. "The Process of International Adjustment under Conditions of Full Employment: A Keynesian View." Speech, Econometric Society, December.

Obstfeld, Maurice. 2024. "Misconceptions about US Trade Deficits Muddy the Economic Policy Debate." Peterson Institute for International Economics Policy Brief 24-7.

Obstfeld, Maurice. 2026. "Global Imbalances Redux." In Paris Report 4: The New Global Imbalances, edited by Hélène Rey, Beatrice Weder di Mauro, and Jeromin Zettelmeyer, 85–121. CEPR Press.

Obstfeld, Maurice, and Kenneth Rogoff. 1995. "The Intertemporal Approach to the Current Account." In *Handbook of International Economics*, Vol. 3, edited by Gene M. Grossman and Kenneth Rogoff, 1731–99. North-Holland.

Obstfeld, Maurice, and Kenneth Rogoff. 2001. "Perspectives on OECD Economic Integration: Implications for U.S. Current Account Adjustment." In *Global Economic Integration: Opportunities and Challenges*, 169–208. Books for Business.

OECD. 2025. "How Governments Back the Largest Manufacturing Firms." OECD Trade Policy Paper 289.

Pettis, Michael. 2024. "Can Trade Intervention Lead to Freer Trade?" Carnegie Endowment for International Peace (Commentary), February 23. https://carnegieendowment.org/china-financial-markets/2024/02/can-trade-intervention-lead-to-freer-trade.

Phillips, Steven, Luis Catão, Luca Ricci, Rudolfs Bems, Mitali Das, Julian Di Giovanni, D. Filiz Unsal, et al. 2013. "The External Balance Assessment (EBA) Methodology." IMF Working Paper 13/272.

Razin, Assaf, and Lars E. O. Svensson. 1983. "Trade Taxes and the Current Account." *Economics Letters* 13 (1): 55–57.

Rey, Hélène, Beatrice Weder di Mauro, and Jeromin Zettelmeyer, eds. 2026. *Paris Report 4: The New Global Imbalances*. CEPR Press.

Roldos, Jorge E. 1991. "Tariffs, Investment and the Current Account." *International Economic Review* 32 (1): 175–94.

Rotunno, Lorenzo, and Michele Ruta. 2024a. "Trade Implications of China's Subsidies." IMF Working Paper 24/180.

Rotunno, Lorenzo, and Michele Ruta. 2024b. "Trade Spillovers of Domestic Subsidies." IMF Working Paper 24/41.

Rotunno, Lorenzo, Michele Ruta, and Priyam Verma. 2026. "Industrial Policy and Trade Tensions in Strategic Sectors." Unpublished.

Schmitt-Grohé, Stephanie, and Martín Uribe. 2025. "Transitory and Permanent Import Tariff Shocks in the United States: An Empirical Investigation." NBER Working Paper 33997.

Sen, Partha, and Stephen J. Turnovsky. 1989. "Tariffs, Capital Accumulation, and the Current Account in a Small Open Economy." *International Economic Review* 30 (4): 811–31.

Smets, Frank, and Rafael Wouters. 2007. "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach." *American Economic Review* 97 (3): 586–606.

Stockman, Alan C., and Linda L. Tesar. 1995. "Tastes and Technology in a Two-Country Model of the Business Cycle: Explaining International Comovements." *American Economic Review* 85 (1): 168–85.

Teti, Feodora. 2024. "Missing Tariffs." CESifo Working Paper 11590.

Yellen, Janet. 2024. "Remarks by Secretary of the Treasury Janet L. Yellen at a Press Conference in Beijing, the People's Republic of China." US Department of the Treasury (press release), April 8. https://home.treasury.gov/news/press-releases/jy2241.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 171–192

Evaluating the Fiscal and Distributional Implications of Tariffs

Kyle Pomerleau and Erica York

A straightforward and uncontested result of tariffs is that they increase government revenue. Prior to President Trump's first term starting in 2017, customs duties only accounted for 1 percent of federal revenue and collected about 0.2 percent of gross domestic product (GDP). However, if tariffs roughly the size of those enacted in 2025 become an ongoing feature of the US budget, then tariffs could become 5 percent or more of federal revenue in the next few years, collecting nearly as much as the corporate income tax.

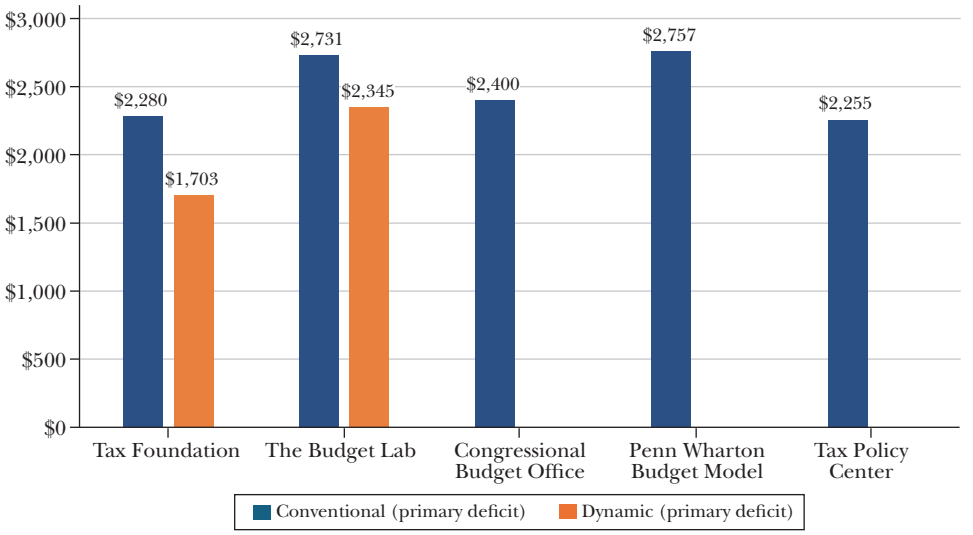
Given the growing importance of tariffs as fiscal policy, it is important to analyze them the same way as other taxes. Tax and spending proposals are "scored" by both governmental and nongovernmental organizations to determine how they will affect federal revenue, spending, debts, deficits, and the distribution of the tax burden. The methodologies for most tax policies are well-established. In contrast, tariffs have not been analyzed at the same depth until very recently. Tariffs are a form of indirect or excise tax, and so many of the current methods used to estimate the revenue implications of excise taxes apply to tariffs. However, tariffs pose their own methodological challenges and conceptual differences.

This paper describes and reviews the approaches used to estimate the revenue and distributional consequences of tariffs. We rely on the revenue estimates from five prominent organizations that regularly produce revenue estimates of current

■ *Kyle Pomerleau is a Senior Fellow, American Enterprise Institute, Washington, DC. Erica York is Vice President of Federal Policy, Tax Foundation, Washington, DC. The authors' email addresses are kyle.pomerleau@aei.org and eyork@taxfoundation.org.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20251490.

Figure 1
Conventional and Dynamic Tariff Revenue Estimates, 2026–2035 Budget Window, in Billions



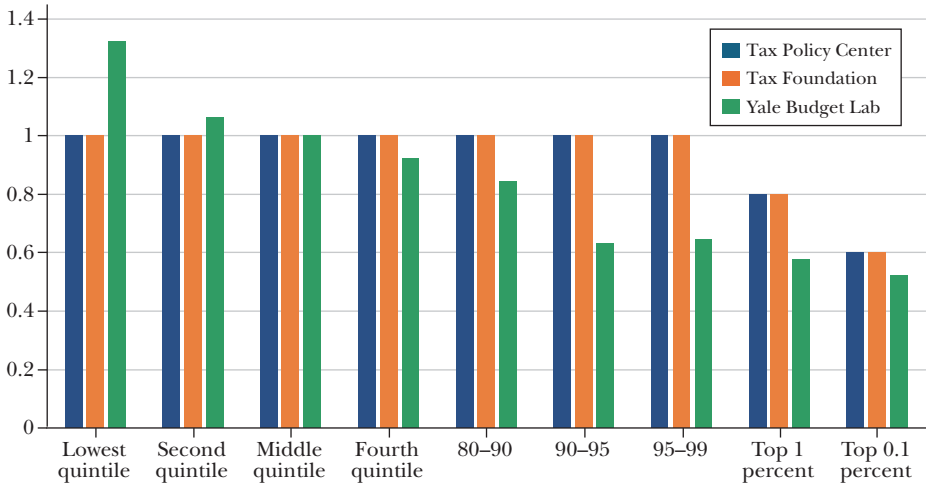
Source: York and Durante (2026), Yale Budget Lab (2026), CBO (2026a), Tax Policy Center (2026), and Penn Wharton Budget Model (2026).
Note: Groups vary slightly in the tariff policies included in their estimates.

tariff policies: the Congressional Budget Office (CBO),[1] the Penn Wharton Budget Model, the Tax Foundation, the Tax Policy Center, and the Budget Lab at Yale.[2] We also rely on the distributional analysis of three of these five groups: the Tax Foundation, the Tax Policy Center, and the Budget Lab at Yale. The organizations find that tariffs in place prior to February 20, 2026, would raise between $2.2 trillion to $2.8 trillion over ten years, conventionally estimated, as shown in Figure 1. Two organizations also produce dynamic revenue estimates, which account for changes in economic growth, and find tariffs would raise slightly less.

Three organizations produce distributional estimates and find that tariffs are regressive: the tax burden as a share of (pre-tariff) after-tax income is higher for low- to middle-income households than for high-income households. Figure 2 shows the distributional pattern of tariff policies as a percent of after-tax income, as estimated by the Tax Policy Center, the Tax Foundation, and the Budget Lab at Yale.[3] Values show tariff burden as a percent of after-tax income for each income group, normalized so that the middle quintile equals 1.0. Values below 1.0 indicate a burden less

[1] The Congressional Joint Committee on Taxation does not produce revenue estimates for tariffs, because tariffs are not part of the Internal Revenue tax code.
[2] For the Budget Lab at Yale, we describe the methodology for the numbers included in this report, but that methodology is in the process of being updated.
[3] Quintile and detailed high-income percentiles not available. Imputed based on data from Tax Policy Center and Tax-Data.

Figure 2
Normalized Distributional Impact of Tariffs as a Share of After-Tax Income (Middle Quintile = 1)



Source: Yale Budget Lab (2025), Tax Foundation (2026), Tax Policy Center (2025b), and authors' calculations.
Note: Values show tariff burden as a percent of after-tax income for each income group, normalized so that the middle quintile equals 1.0. Values below 1.0 indicate a burden less than the middle quintile and values above 1.0 indicate a burden higher than the middle quintile.

than the middle quintile and values above 1.0 indicate a burden higher than the middle quintile. This approach allows comparison of the relative burden when the overall burden (size of the tax increase) differs.

The bulk of the tariffs underlying these estimates were ruled unconstitutional in the US Supreme Court case of *Learning Resources, Inc. vs. Trump* (Docket #24-1287, February 20, 2026), but then were rapidly replaced with an alternative set of tariffs with a different legal justification. By the time this paper is published, or soon after, yet another new set of tariffs may well be in place. Thus, our focus in this essay will place little emphasis on specific revenue estimates or distributional analyses but instead use the 2025 tariffs as a running example to explain the data and methods behind the tariff scores.

Scoring Basics

Revenue analyses follow certain conventions. Conventional revenue estimates hold nominal gross domestic product (or gross national product) constant. However, taxpayers can have behavioral responses to taxes that affect revenue implications of a policy (Joint Committee on Taxation 2025b). A classic example is that taxing cigarettes reduces quantity demanded, lowering the revenue the tax

would otherwise raise. Organizations also produce "dynamic" revenue estimates. These scores capture both behavioral responses to tax policy and how these taxes affect economic output or national income and thus alter the estimates of how the amount a given tax change impacts federal revenue (Swagel and Bartold 2025).

Distributional analyses of tariffs provide information on how the policies will burden households, most frequently by how changes in taxes impact a household's after-tax resources by income group. Most groups rank households by an "expanded" definition of income, which includes market income (wages, salaries, fringe benefits, and capital income) and government benefits. Similarly, standard distributional analyses hold total output fixed.

Revenue Estimates

Organizations vary in how they estimate the revenue and budget effects of tariffs. This includes choices about the data and methods to determine taxable imports; behavioral reactions to tariffs; the possibility of noncompliance with tariffs; how to calculate gross tariff revenue; the "excise tax offset," which is the extent to which income and payroll tax revenue falls in response to higher excise taxes; and whether to account for how tariffs impact the value of government spending. Table 1 summarizes the methodological choices used by five tariff-scoring organizations.

Matching Taxable Imports and Tariffs

The starting point for all estimates of tariff revenues is an estimate of taxable imports. This requires matching the value of imports subject to the applicable tariff rate. Official import data may be accessed through the US Census Bureau (2025) or the US International Trade Commission (2025) DataWeb tool, which match import values with tariff rates. Both datasets allow users to classify import data by Harmonized Tariff Schedule (HTS) codes. These codes are required to match import values to the tariff rates published in official tariff announcements published in the Federal Register.

The Congressional Budget Office, the Tax Policy Center, the Tax Foundation, and the Penn Wharton Budget Model all utilize data from the US Census Bureau on imports, while the Budget Lab at Yale utilizes the Global Trade Analysis Project (GTAP) model, coordinated by the Center for Global Trade Analysis at Purdue University, which contains import data from the UN Comtrade database and must map the GTAP sectors to the official Harmonized System codes.

Matching tariff policies with imports poses at least three challenges. First, tariff rates may interact or "stack." For example, one type of imported good may be subject to multiple cumulative tariffs that stack up; in other cases, only one specified tariff may apply in cases where a good could be subject to multiple levies (Whitehouse 2025).

Second, some tariffs may be applied to goods based on characteristics for which no clear data source exists: for example, derivative goods made using steel and aluminum subject to the Section 232 tariffs or pharmaceutical tariffs based on

Table 1
Tariff Revenue Estimates and Methodological Choices

	Tax Policy Center	Tax Foundation	Congressional Budget Office	The Budget Lab at Yale	Penn Wharton Budget Model
Date/policies	Tariffs announced from January 20, 2025, through January 8, 2026	Tariffs implemented and scheduled as of February 6, 2026	Actions taken through November 20, 2025	Tariffs implemented on January 19, 2026	Tariffs announced as of November 13, excluding temporary or speculative changes
Conventional score, 2025–2036 (revenue raised)	$2,255 billion	$2,280 billion	$2.4 trillion, plus an additional $0.5 trillion reduction in interest costs	$2,731 billion	$2,757 billion
Elasticity	Varies by product, weighted sum −4.2	−2	Nested constant elasticity of substitution (CES) model. Elasticities vary by industry by center on 5 for foreign-to-foreign substitution, 3.3 within-industry substitution, and 2.5 foreign-to-domestic substitution	−1	Calibrated to −2.12, but varies with tariff rate
Compliance	No additional adjustment	8 percent noncompliance	No additional adjustment	10 percent noncompliance	No additional adjustment
Gross tariff revenue	Post-behavior import values times the statutory tariff rate	Post-behavior import values times the tax-inclusive statutory tariff rate	Post-behavior import values times the statutory tariff rate	CBO projections of customs duties grossed up for changes in the effective statutory tariff rate on imports.	Post-behavior import values times the statutory tariff rate
Income and payroll offset	JCT's Offsets. 0.244 in 2025	Tax Foundation Microsimulation Offsets 0.23 to 0.24	JCT's Offsets. 0.244 in 2025	0.23 assumption	JCT's Offsets. 0.244 in 2025
Other budgetary effects	None	None	None	None	None

Source: York and Durante (2026), Yale Budget Lab (2026), CBO (2026a), Tax Policy Center (2026), and Penn Wharton Budget Model (2026), plus additional information from corresponding with individuals from each organization through a series of meetings and emails.
Note: For the Budget Lab at Yale, we describe the methodology for the numbers included in this report, but that methodology is in the process of being updated.

company-level exemptions. No database exists to identify the metal content of all derivative goods, introducing uncertainty into estimating the taxable import base. This problem may also interact with the stacking challenge, as the nonmetal content may be subject to other tariffs.

A third challenge is that certain imports under the United States-Mexico-Canada Agreement (USMCA)—which replaced the earlier North American Free Trade Agreement (NAFTA) and took effect in 2020—from Canada and Mexico are not subject to the new tariffs. Prior to 2025, compliance with USMCA rules was lower because the tariff differential was small. In 2024, 38 percent of imports from Canada and 49 percent of imports from Mexico officially qualified for USMCA treatment. After new tariffs, by August 2025, 86 percent of Canadian imports and 87 percent of Mexican imports qualified for tariff relief under USMCA, significantly shrinking the tax base. Using 2024 USMCA import shares would suggest more than $253 billion of goods imports from Mexico would have faced tariffs on an annual basis; using more recent shares shrinks this to approximately $55 billion in taxable imports on an annual basis.

Behavioral Responses and Noncompliance

Tariffs raise the relative price of imported goods. This relative price effect creates behavioral responses that have a direct effect on the amount of revenue that tariffs raise. Import flows can change along several margins, with varying magnitudes along those margin: for example, consumers switching between taxed imports and untaxed domestic goods and switching between taxed imports and untaxed imports. In addition, importers and exporters may cooperate to avoid tariffs by rerouting trade through third countries that are not subject to the same taxes. These behavioral responses are captured in trade elasticities, which measure the percent change in the quantity of imports given a percent change in the price of imports.

Gravity-based trade model methods are among the most widely used tools for estimating these elasticities. They exploit variation in trade costs such as tariffs to identify how trade flows respond to price changes (Head and Mayer 2014). A key elasticity in the trade policy literature is the so-called Armington elasticity, which measures the substitutability between goods from different countries. At a macro level, it measures the elasticity of substitution between domestically produced goods and imports, and at a micro level, it measures the elasticity of substitution between imports from different countries. A higher Armington elasticity means buyers readily switch between goods from different sources, while a lower elasticity means they do not easily substitute.

The substitutability between domestic production and imports is generally estimated to be lower than the substitutability between different import suppliers (Feenstra, Luck, Obstfeld, and Russ 2018). The literature also suggests that long-run elasticities are greater than short-run elasticities. For example, in the United States in 2018 and 2019, a 10 percent tariff was associated with a 10 percent drop in imports over the first three months and a 20 percent drop over the following months (Amiti, Redding, and Weinstein 2019). A review of the 2018–2019 tariffs on China by the US International Trade Commission (2023) found an elasticity close to −1 in the first few months of the tariffs that rose to more than −2 near the end of the second year.

The behavior modeled and elasticities employed by organizations vary significantly. The Congressional Budget Office utilizes a nested constant elasticity of

substitution (CES) model in Fajgelbaum, Goldberg, Kennedy, and Khandelwal (2020). In this approach, imports respond to the relative price of imports along three margins each with a different elasticity. First, there is an incentive to substitute purchases from one country to another to avoid levies on specific goods from specific countries (foreign-to-foreign substitutions). For example, a consumer may purchase Vietnamese coffee to avoid levies on Colombian coffee. Second, US households and businesses may substitute away from tariffed foreign goods toward foreign substitutes (within-industry substitution). For example, consumers may purchase foreign tea instead on foreign coffee. Finally, US consumers may substitute from foreign tariffed goods to domestic substitutes (foreign-to-domestic substitution) (Fajgelbaum, Goldberg, Kennedy, and Khandelwal 2020). This outcome would occur if US consumers purchased US-produced steel instead of steel from any foreign producer. The elasticities used by the CBO vary by industry, but are centered around 5 for foreign-to-foreign substitution, 3.3 for within-industry substitution, and 2.5 for foreign-to-domestic substitution. For example, a foreign-to-domestic elasticity of 2.5 means a 1 percent rise in the price of foreign steel relative to domestic steel reduces the ratio of foreign-to-domestic steel purchased by 2.5 percent.

The Tax Policy Center, in contrast, uses an elasticity of import demand with respect to import prices. Their elasticities vary by industry but have a weighted sum of −4.2 across all goods. The Budget Lab at Yale utilizes the Global Trade Analysis Project model to produce behavioral responses to tariffs. Those responses vary by product and industry, and the exact elasticities are not well-documented, but overall the responsiveness of imports to changes in tariff rates in the GTAP implies an elasticity across all products of −1. The Tax Foundation uses an elasticity of −2, but it does not vary across products or industries. The Penn Wharton Budget Model is calibrated to a long-run elasticity of −2.12, but the actual elasticity from the model is a function of the tariff rate and includes multiple elasticities (foreign-to-foreign and foreign-to-domestic, which in practice typically results in a higher elasticity) (based on Boehm, Levchenko, and Pandalai-Nayar 2023).

All groups except for the Tax Foundation assume that the behavioral response phases in over time. The time path utilized by those groups (except the Budget Lab at Yale) is consistent with Boehm, Levchenko, and Pandalai-Nayar (2023). As such, the elasticity is roughly one-third its long-run value in the first year and rises linearly to its long run value after seven years. For example, the Tax Policy Center's weighted average elasticity in the first year is −1.6 and grows to −4.2. The Congressional Budget Office reaches its long-run values over roughly ten years. The Budget Lab at Yale imposes a time path for behavioral responses in their revenue score; in GTAP, shifts in import weights are assumed to occur immediately, while total import adjustments are assumed to phase in linearly over three years.

The groups use a wide range of methodologies for estimating how taxable imports change in response to tariffs. While the overall magnitude of behavioral responses does not seem to vary significantly for the 2025 tariff policies, other policies may result in larger differences. For example, the methods used by the Congressional Budget Office and the Tax Policy Center, which capture variation

across products and industries, may show larger or smaller behavioral responses in some cases than groups that use a single elasticity.

Applying the literature of trade elasticities to the recent increase in US tariff policies has two potential limitations. First, the magnitude of US tariffs starting in April 2025 is significantly higher than those studied in the literature, and the uncertainty and variability may also have different effects. Responsiveness may be nonlinear and could either rise or decline at much higher tax rates. Second, the tariffs being utilized are much broader based than tariffs in recent history. Imports may be more responsive to tariffs when they narrowly apply to certain products, countries, or industries—because there is more opportunity to avoid taxation. When tariffs apply to all countries and products, the ability to avoid tariffs will decline along some margins.

Some organizations make additional adjustment for "noncompliance," which includes behaviors such as rerouting shipments through third countries to misrepresent origin, as well as undervaluation of customs declarations. The Tax Foundation and The Budget Lab at Yale assume that a fixed amount of the tariff base is lost due to noncompliance, 8 percent and 10 percent, respectively. This "haircut" is fixed regardless of how high tariff rates are.

The choice of modeling additional compliance has a direct effect on estimated revenue collections. Those that choose to model noncompliance separately find less overall revenue than those that do not. One potential justification for a noncompliance adjustment in addition to application of trade elasticities is that the tariff policies the trade elasticity literature are based on are, in general, not as high and variable as the 2025 tariffs. Given the larger increase in tariffs, noncompliance responses could be larger than what is captured in the trade elasticity literature.

Gross Tariff Revenue

After determining taxable imports (after behavior), groups then calculate gross tariff revenue. Gross tariff revenue is the amount of revenue raised directly from the tariffs. Most groups have a similar method for calculating gross tariff revenue. The Congressional Budget Office, Penn Wharton Budget Model, and Tax Policy Center calculate gross tariff revenue by multiplying the statutory tariff rate by the post-behavioral response value of imports. The Tax Foundation, in contrast, calculates gross tariff revenue by multiplying post-behavior imports by the tax-inclusive tariff rate (the statutory tax rate divided by 1 plus the statutory tax rate).[4]

Although this step may seem straightforward, estimators need to take care under a fixed GDP (or GNP) setting (Kitchen 2017). Excise taxes create a wedge between the prices that consumers pay and the prices that producers receive. For a given good, the extent to which the tax is "passed" to a consumer in the form of higher prices or a producer in the form of lower income depends on the relative responsiveness of the two parties. As a result, if the price of one good rises due to

[4] This is consistent with how the Tax Policy Center has estimated gross value-added tax (VAT) revenue in the past (see Toder et al. 2012a).

an excise tax, the price of all other goods and producer prices must fall. Keeping track of these relative price effects is important in estimating the ultimate budgetary effects of tariffs (Gale and Pomerleau 2023a).

If estimators are not careful, they can either overstate or understate the revenue effects of excise taxes like tariffs. For example, multiplying taxable imports by the statutory tax rate can be consistent with the assumption that the Federal Reserve accommodates the tax and that prices rise by the full amount of tariffs while producer prices (the amount of income that workers and owners of capital receive) remain constant (Kitchen 2017). Without additional adjustments, this would be inconsistent with holding nominal GDP fixed and overstate the real revenue effects of tariffs because the higher price level means that any given dollar of revenue and federal spending can purchase fewer goods and services.[5] No organization has explicitly modeled how their estimates hold nominal GDP fixed.

Income and Payroll Tax Offsets

Another important implication of tariffs is their impact on income and payroll tax revenue. Tariffs reduce real income and payroll tax revenue through what is called the "excise tax offset." The excise tax offset occurs because, under a fixed GDP assumption tariffs mechanically reduce the amount of nominal income received by workers and owners of capital: the amount the federal government collects in excise tax revenue is unavailable to for the federal government to tax as wages or profits.[6]

All organizations account for the reduction in income and payroll tax revenue or the "excise tax offset." The Congressional Budget Office, Penn Wharton Budget Model, and Tax Policy Center all utilize the Joint Committee on Taxation (2025) projection of the excise tax offset, which varies by year and is 0.244 in 2025 and 0.261 in 2035.[7] The Tax Foundation uses an excise tax offset produced with their microsimulation model and it ranges from 0.23 to 0.24 over the next decade. The Budget Lab at Yale assumes a flat 23 percent offset over the entire decade.

Tariffs add a potential complication to the standard income and payroll tax offset because consumers and producers are located in two different taxing jurisdictions. If a tariff is passed to a foreign producer, that would reduce foreign income

[5] For example, if prices rose by 5 percent to accommodate a 5 percent excise tax and nominal GDP rose from $100 to $105 ($5 being the amount the government collects), the government would only be able to purchase $4.76 expressed in pre-tax prices.

[6] The offset occurs regardless of who bears the initial burden of the excise tax. On one extreme, a taxed good could have a perfectly inelastic supply and perfectly elastic demand. In this case, the firm is unable to pass the tax on to consumers and bears the full burden. This reduces the amount of income available to compensate workers and owners of capital, resulting in a reduction in income and payroll taxes. On the other extreme, demand is perfectly inelastic and supply is perfectly elastic. In this case, firms are able to fully shift the entire burden of the excise tax on to consumers. Holding demand constant, consumers have less money to spend on non-taxed goods. This reduces revenues in other industries, reducing the amount of income and payroll taxes collected (Joint Committee on Taxation 2011).

[7] These values were calculated before the passage of the One, Big, Beautiful Bill Act signed into law in July 2025, so they are subject to change.

and payroll tax revenues, not American tax revenues. Assuming no changes in exports, $1 of tariff revenue could raise more than an equivalent domestic excise tax. However, research on tariffs imposed during the first Trump administration suggests that US consumers bore close to the entire burden of tariffs (in this journal, Amiti et al. 2019; Neiman and Gopinath in this symposium). Even in cases where incidence falls of foreigners, other mechanical effects have resulted in domestic prices rising by the amount of the tariff (Ganapati and Hottman 2026). As a result, it is appropriate to assume the entire excise tax offset.

Additional Budgetary Effects of Tariffs

In addition to revenue effects, tariffs can have an impact on the value of federal spending. Excise taxes, such as tariffs, can apply to government purchases. When those government purchases are taxed, the government raises more revenue, but spending must rise by the same amount to cover the new tax. As a result, taxing government spending has no net effect on the government budget (Gale and Pomerleau 2023a). When nominal GDP is held fixed and government spending is exempt, nominal government spending can fall in line with producer prices and hold real federal purchases fixed (Toder, Nunns, and Rosenberg 2011). In contrast, if government spending is subject to tax, nominal spending can be held fixed to hold real spending constant. The federal government may be exempt from customs on imports, but it can still be indirectly burdened if it purchases goods from suppliers that pass along tariffs as higher prices.[8]

Currently, no group explicitly models the full budgetary effects of tariffs. That said, the net budgetary effects of tariffs are unlikely to deviate significantly from current revenue estimates.

Other Methodological Differences

In addition, organizations differ in how they project future imports. The Congressional Budget Office produces projections of total imports over the next decade. The other groups we consider build upon these projections, except for the Penn Wharton Budget Model. In its latest baseline, which incorporates a behavioral response to the tariff policy assumed in the baseline, the CBO projects nominal imports will total $48.2 trillion from 2026 through 2035. The Penn Wharton Budget Model projects under current tariff policy, which reflects lower tariff rates than the CBO baseline, slightly more than $39 trillion in imports for the same budget window. As a result, the Penn Wharton Budget Model projects slower growth of taxable imports than all other groups.

[8] The Acquisition.gov website of the US government states: "United States laws impose duties on foreign supplies imported into the customs territory of the United States. Certain exemptions from these duties are available to Government agencies. Agencies must use these exemptions when the anticipated savings to appropriated funds will outweigh the administrative costs associated with processing required documentation" (see https://www.acquisition.gov/far/subpart-25.9#:~:text=United%20States%20laws%20impose%20duties,associated%20with%20processing%20required%20documentation).

Projecting tariff revenues into the future also requires decisions about how to treat tariffs that are announced but not yet enacted, or enacted but are scheduled to phase out. The Congressional Budget Office models "changes announced by the Administration after they are implemented," which excludes tariffs that have been announced or scheduled but not implemented, while treating tariffs in effect as permanent (Swagel 2025c). The Yale Budget Lab (2025) follows the same guidelines. In contrast, the Tax Foundation models tariff policy as announced, including temporary pauses and scheduled increases or repeals in tariff rates. The Tax Policy Center (2025a) also models tariffs as announced, and the Penn Wharton Budget Model (2026) models a scenario of "all tariffs announced by the White House" as of a certain date in their estimates, but excludes "temporary suspensions, short-term pauses, and speculative changes."

Variations in these assumptions can lead to meaningful variation in revenue estimates. Modeling tariffs as they are currently in effect captures the revenue that would be expected if current policy continued without changes. Modeling tariffs as they are in effect and announced changes captures the revenue that would be expected from current and planned policy.

Dynamic Revenue Estimates

Dynamic revenue estimates generally start with the conventional revenue estimates. Then changes in total economic output are estimated, which induce changes in taxable income, which further influence federal revenues. The dynamic revenue effect of tariffs is negative—tariffs raise less revenue when considering their economic effects. First, in the short run, tariffs reduce real after-tax household income, which reduces the demand for goods and services and the demand for labor and capital, thus reducing economic output in the short run. Second, in the long run, tariffs reduce economic output by reducing returns to work and investment in the United States. This will lead to fewer hours worked, lower productivity, and lower wages. Third, tariffs raise net revenue, which reduces the amount of federal borrowing and "crowds in" private investment, which increases long run output.

As noted earlier, two groups have produced dynamic revenue analyses of current tariffs, summarized by Table 2.[9] The Yale Budget Lab (2026) finds the tariffs would slow output growth by 0.4 percentage points in 2026 and reduce long-run output by 0.3 percent. The Tax Foundation finds the tariffs would reduce long-run output by 0.5 percent (York and Durante 2026). The reduction in economic output results in lower revenue collections when dynamically scored (as shown earlier in Figure 1). Over a ten-year horizon, the Yale Budget Lab estimates incorporating macroeconomic feedback causes the revenue to fall by about 14 percent, while the

[9] The Penn Wharton Budget Model has produced dynamic analyses of past tariff policies (see Boller et al. 2025).

Table 2
Dynamic Estimates of Tariff Revenues

	Tax Foundation	The Budget Lab at Yale
Budget window	2026–2035	2026–2035
Tariff policy	Tariffs implemented and scheduled as of February 6, 2026	Tariffs implemented on January 19, 2026
Conventional revenue (primary deficit)	$2,280 billion	$2,731 billion
Dynamic revenue	$1,703 billion	$2,345 billion
Dynamic revenue feedback	−$577 billion	−$386 billion
Long-run GDP effect	−0.5 percent	−0.3 percent

Source: York and Durante (2026) and Yale Budget Lab (2026).

Tax Foundation estimates incorporating macroeconomic feedback causes tariff revenue to fall by about 25 percent.

Distributional Implications of Tariffs

Tariffs, as a form of excise tax, reduce after-tax incomes by either reducing the income received by households or by raising the prices that households pay for goods and services. The distribution of the burden depends on the types of goods and services subject to taxation, the distribution of different forms of household income, and differences in consumption patterns across households. Table 3 summarizes the choices made behind these calculations. Methods vary in the data they use, how households are classified, and how tariffs are distributed to households. There are several additional distributional implications that no group currently addresses and that we discuss below.

Differences in Data and Definitions

Distributional analyses are typically produced using microsimulation models. These models have preprogrammed rules that can simulate the impact of different policies on a representative sample of US households or US tax filers. Most often, the core data used for tax-based models are the Internal Revenue Service Public Use File (PUF), which is a stratified random sample of roughly 150,000 tax filers in the United States that represents the 160 million tax filers in the United States. The data include information from these tax filers' tax returns such as total income, deductions, credits, and some other information necessary to compute tax liabilities. However, it does not include demographic or spending pattern information of households, individuals or households that do not file taxes, or sources of income not reported to the IRS (such as government benefits and increases in net worth through appreciation of stock). To address these shortcomings, organizations typically impute this information from other sources.

Table 3
Data and Definitions for Distributional Analysis of Tariffs

		Tax Policy Center	Tax Foundation	The Budget Lab at Yale
Data		IRS Public Use File, augmented with information from Current Population Survey, Survey of Consumer Finances, and Consumer Expenditure Survey	IRS PUF, Augmented with information from Current Population Survey	Consumer Expenditure Survey (CEX)
Income classifier		"Expanded Cash Income"	"Expanded Income"	CBO's Post Tax, Post Transfer Income
How tariffs are distributed	"Sources"	Household burden based on share of Labor Compensation, Wage Indexed government benefits, and Above-normal returns to capital	Household burden based on share of Labor Compensation and Above-normal returns to capital	N/A
	"Uses"	Captures relative price effects using an input-output model.	N/A	Household burden based on ratio of consumption-to-household income and share of consumption devoted to taxed goods.
Transitional burden		None; Long run	None; Long run	N/A
Excise tax offset		Explicitly modeled	Explicitly modeled	Explicitly modeled

Source: Yale Budget Lab (2025), Tax Policy Center (2022), and Li, Watson, and York (2025).

The Tax Foundation and Tax Policy Center have microsimulation models that utilize an augmented IRS Public Use File.[10] The Tax Foundation uses the open source "Tax-Data" program to impute additional household demographic information to tax returns from the Current Population Survey (Li, Watson, and York 2025). The Tax Policy Center (2022) uses a similar imputation method, but also imputes further information on household spending, wealth, and demographics from data sets that include the Survey of Consumer Finances and the Consumer Expenditure Survey. The Yale Budget Lab's (2025) model, in contrast, uses a microsimulation model based on the Consumer Expenditure Survey data.

To analyze how taxes affect households at different income levels, organizations need to rank households by a measure of household well-being based on income. Generally, organizations use what are referred to as "expanded" definitions of income that capture household wellbeing. For example, the Tax Policy Center (2022) utilizes the most comprehensive measure of household income, called "Expanded Cash Income," which is the sum of adjusted gross income on tax returns, deductions allowed in the tax code, employee and employer (including employer-side payroll taxes) contributions to tax-preferred retirement accounts, pension income, government benefits, and the household's share of the corporate income tax.

[10] The Tax Policy Center uses the 2006 PUF and the Tax Foundation uses the 2011 PUF. Both groups "age" the files to current values using aggregate targets from IRS and Congressional Budget Office data.

A relatively broad definition of income has some advantages. For example, tariffs are imposed on goods that are purchased with government benefits. If tariffs are imputed to households based on the goods purchased as shown in the Consumer Expenditure Survey, but the measure of income does not include government benefits, then the burden of tariffs will be misstated.

Allocating the Burden of Tariffs

Conceptually, an excise tax burdens a household in two ways. First, an excise tax reduces the value of certain after-tax "sources" of household income. For example, if all goods and services are 10 percent more expensive due to an excise tax, the real after-tax value of income will fall by 9.09 percent.[11] Second, excise taxes burden the "uses" of household income by raising the price of goods and services. Households that devote a larger share of their household income to the consumption of taxed goods will face a higher tax burden than those who devote a smaller share of their consumption to those goods.

If an excise tax applies equally to all goods and services purchased by households, the sources burden of an excise tax is equal to the uses burden in present value (Toder, Nunns, and Rosenberg 2012). The sources and uses of income burdened by an excise tax can be demonstrated with an identity (Tax Policy Center 2011). A household's total after-tax income is equal to a household's total uses of income for consumption of saving:

$$\textit{Labor Compensation} + \textit{Capital Income} + \textit{Government Benefits} - \textit{Tax}$$

$$= \textit{Consumption} + \textit{Saving}.$$

Labor compensation includes salaries, wages, and fringe benefits. Capital income is net business income, interest, dividends, and capital gains. Households can also receive income in the form of government benefits such as Temporary Assistance for Needy Families (TANF) and Social Security benefits or in-kind transfers such as Supplemental Nutritional Assistance Program (SNAP) and Medicaid.

A tax (t) that applies to consumption on the right-hand side would fall on labor compensation, capital income (net of saving), and government benefits (net of tax):[12]

$$(\textit{Labor Compensation} + \textit{Capital Income} - \textit{Saving} + \textit{Government Benefits} - \textit{Tax})/(1 + t)$$

$$= \textit{Consumption} \times (1 + t).$$

[11] A 10 percent tax-exclusive excise tax raises the price of a \$1.00 good from \$1.00 to \$1.10. A consumer with \$1.00 of income can now only purchase $1/1.10 = 0.909$ units of the good, a decrease of 9.09 percent. Generally, a tax-exclusive rate of t reduces real after-tax income by $t/(1 + t)$.

[12] As discussed below, the burden on government benefits depends on whether and how government benefits are adjusted in response to the imposition of an excise tax.

Labor compensation is burdened whether or not it is saved, because households face the same tax burden on their labor income whether they spend their income today, or save it, earn a return that compensates them for waiting, and spend it in the future (Gale and Pomerleau 2023b). Capital income is only subject to tax to the extent to which it exceeds the "normal" return to saving, because the implicit deduction for saving offsets that return in present value.

For tariffs, which do not apply to all goods, the uses-based burden will depend on how households allocate their consumption (the right-hand side of the equation) to the products that face tariffs. Households that devote a larger share of household income to purchasing taxable imports face a larger tax burden than households that do not. For example, some research suggests that low-income households allocate a larger share of their consumption to tradeable goods, which implies tariffs can be more regressive than a broad-based tax on all goods and services (Fajgelbaum and Khandelwal 2016). However, other research suggests that the share of consumption devoted to imported goods is relatively stable across the income distribution or might rise (Atkin 2024).

Tariffs, however, do not only apply to household consumption. Tariffs also apply to imported capital and intermediate goods purchased by businesses. Tariffs that fall on investment goods can burden the normal return to capital (Pomerleau and York 2025; Viard 2010; Joulfaian and Mackie 1992). In addition, the burden placed on the normal returns can also distort investment, which can shift the burden of tariffs onto other taxpayers. In the case of corporate taxation, it is standard to assume that the burden a tax places on the normal return is split between owners and workers. Owners of capital (households that receive returns on corporate equities, debt, and noncorporate businesses) would experience lower after-tax return. Workers would experience lower labor compensation due to lower investment and labor productivity (Nunns 2012). About 20 percent of imports are industrial supplies and materials and another 29.7 percent are capital goods (excluding automobiles), according to the US Bureau of Economic Analysis. As a result, tariffs more broadly apply to all sources of income (labor and capital income) and their uses (consumption and saving). Notably, the burden tariffs place on capital income cannot be avoided by saving because capital goods are subject to tax.

All groups that produce distributional analyses treat tariffs as excise taxes that fall exclusively on household consumption. However, groups vary to the extent to which they account for how tariffs burden the sources of income and the uses of that income.

The Tax Policy Center's methodology for estimating distributional effects of tariffs captures both sources- and uses-effects of tariffs. Their tariff burdens start with an allocation based on sources of household income. As such, each household's burden is equal to its share of labor compensation, above normal returns to capital, and government benefits.[13] After allocating the burden based on a household's

[13]The Tax Policy Center assumes that government benefits are indexed by wages in the long run. As a result, tariffs, which reduce the real after-tax value of wages, will reduce government benefits.

source of income, they incorporate the uses-based burden by capturing the relative price effect of tariffs. Households that consume a greater-than-average share of the taxable goods on which the tariff applies face a higher-than-average burden while households that consume lower-than-average share of taxable goods face a lower-than-average burden, with the net effect across all taxpayers being zero. The Tax Policy Center determines these relative price effects using an input-output model (Tax Policy Center 2026).

The Tax Foundation uses a pure sources-based approach to allocating the burden of tariffs to households. The burden of tariffs on a given household depends entirely on its share of labor compensation and the excess of capital income above saving. The Tax Foundation does not include government benefits in their microsimulation model (an assumption also utilized by the US Treasury Department), which implicitly assumes that government benefits do not depend on real wages in the long run and are adjusted for any changes in nominal prices.

The Budget Lab distributes the burden of tariffs to income groups based on household consumption expenditures. They start with each income quintile's share of total consumption expenditures estimated from the Consumer Expenditure Survey. The tariff burden for each income group then depends on an estimated price effect of tariff policy. As a result, the burden for each income group depends on that group's consumption expenditures as a share of its after-tax-after-transfer income and the share of consumption devoted to taxed goods.

The most consequential methodological choice is whether to impute the tax burden based on annual consumption or sources of income (Yale Budget Lab versus the Tax Foundation and Tax Policy Center). Distributing the burden based on annual expenditures makes tariffs appear much more regressive. However, it poses some conceptual challenges. In standard distributional tables, all other taxes are allocated to households when that income is earned. Thus, allocating taxes based on when a household spends income, rather when the income is earned, is inconsistent. Furthermore, a household's tax burden under a uses-only method is highly sensitive to the accounting period, because household consumption and income can vary widely year to year. As a result, tax burdens based on current period consumption as a share of current period income can overstate or understate the actual change in wellbeing.

The Tax Foundation and Tax Policy Center may also be slightly overstating the regressivity of tariffs by excluding the burden they place on capital goods. Incorporating this burden would increase the tax on capital income, which is earned disproportionately by high-income households. The Budget Lab's current framework would make incorporating this burden difficult.

Distributing the Excise Tax Offset

Tariffs result in an "excise tax offset" to income and payroll taxes, because tariffs reduce the amount of real after-tax factor income available to compensate workers and owners of capital so that there is less income subject to taxation under the income and payroll tax. The excise tax offset, like the tariffs themselves, represents

a change in tax liability that should be distributed to households. The three organizations that carry out distributional analyses all account for the excise tax offset in their modeling and explicitly model the burden of the excise tax offset using their microsimulation model.[14]

The excise tax offset is regressive, because the net tax burden of individual and payroll taxes is progressive (Carloni and Dinan 2021). For a low-income household, the excise tax offset may actually raise their individual income tax burden. This is because very low-income households benefit from the Earned Income Tax Credit, and a reduction in reported wages will reduce EITC benefits. In contrast, a high-income household may face the top marginal tax rate of 37 percent. For this household, each dollar of reduced wage income would reduce income tax liability by $0.37.

The Foreign Dimension of Distributional Issues

When a tariff is announced (and expected to be permanent), standard theory suggests that there will be pressure for the nation's exchange rate to appreciate immediately. This appreciation shifts a portion of the burden off households who devote a portion of their income towards consuming imports and onto households that earn income from exports (Pomerleau and York 2025). It is not immediately clear how this shift influences the distributional implications of tariffs. If workers and owners of capital in export industries, for example, have higher-than-average incomes than the broader population, it would make tariffs slightly more progressive. However, if export industries have lower-than-average incomes, it would make tariffs slightly more regressive. In addition, a US-dollar appreciation results in a one-time reduction in the value of all US-owned foreign assets, because investors would require a larger amount of foreign currency to receive the same amount of US dollars than before the dollar appreciation. At the same time, foreign investors that hold US assets receive a one-time windfall, because the US dollars they earn on existing wealth converts to a larger amount of foreign currency.

Tariffs can also partially burden the household income of foreign producers and consumers. Foreign producers receive income from selling goods to Americans, and foreign consumers receive goods from US producers. Depending on the relative supply and demand of goods and services, a portion of the burden of US tariffs could be shifted to foreigners. If that is the case, that portion of tariffs should be removed from the distributional analysis because it is not a burden on US households.

However, the distributional effects of tariffs that can happen through exchange rate changes are not explicitly modeled in tariff-scoring. One reason is that although theory suggests that the higher US tariffs should lead to an exchange rate appreciation, the US exchange rate did not in fact appreciate in 2025, likely because policy uncertainty overset the standard response (Kalemli-Özcan, Soylu, and Yildirim 2026). In addition, most research suggests that tariffs, at least initially, result in higher import prices passed through to US buyers (Amiti, Redding, and Weinstein

[14]The Budget Lab at Yale has a separate individual income and payroll tax microsimulation model that they use to produce their excise tax offset.

2019; Cavallo, Gopinath, Neiman, and Tang 2021). This evidence implies that little to none of the burden of higher tariffs is shifted to foreign households.

Distributional Effects on Wealth and Government Benefits

Tariffs place an additional tax burden on existing wealth. In the long run, taxes on household consumption do not burden the normal return to capital, because taxpayers receive an implicit deduction for saving. However, consumption-based taxes do burden the returns on wealth accumulated before the enactment of the tax (Toder, Nunns, and Rosenberg 2011).

Tariffs can also have a different impact on government benefits in the short run than in the long run. As one example, initial Social Security benefits are determined by the wages that a retiree earns during their life. Tariffs reduce the real after-tax value of wages, and thus directly reduce the value of initial benefits for new retirees. As noted above, the Tax Policy Center assumes that most government benefits have this quality. However, if tariffs result in price increases, benefits are protected through automatic cost of living adjustments. If tariffs do not increase prices, but reduce factor income, government benefits are still protected because they maintain their purchasing power at their original nominal value.

Both the Tax Policy Center and the Tax Foundation distributional analyses seek to measure the long-run burden of tariffs. Therefore, neither group measures the short-run implication tariffs have on holdings of wealth. Furthermore, the Tax Policy Center assumes that government benefits are burdened by tariffs. As discussed above, the Tax Foundation does not include government benefits in their analysis. This approach results in a burden of tariffs on benefits that is more similar to a short-run analysis, as it implicitly assumes that government benefits do not face taxation.

The Yale Budget Lab describes its distributional burden as a short-run burden of tariffs and captures some transitional implications for wealth. For example, a retiree may be consuming and paying tariffs out of existing wealth (however, they may not be capturing that income when measuring the household's wellbeing). Its approach also assumes that tariffs increase the price level and increase transfer payments through cost-of-living adjustments. However, its methodology does not explicitly distinguish between returns on new and existing wealth, nor does it capture the burden that tariffs place on wealth that is not consumed within the accounting period.[15]

Given that wealth is highly concentrated among high-income households, the burden tariffs place on existing wealth in the short run is highly progressive. Likewise, the fact the tariffs place little burden on the value of government benefits in the short run makes tariffs more progressive in the short run than the long run.

[15] One could justify a method similar to the Budget Lab's by arguing that if tariffs are temporary, taxpayers could genuinely avoid tariffs by saving. However, this is inconsistent with the assumptions that the Budget Lab makes when it estimates the economic and budgetary effects of tariffs.

Conclusion

After decades of near irrelevance to federal revenues, tariffs now raise a meaningful amount of federal revenue. Policymakers have a keen interest in how taxes impact federal revenues and the distribution of the tax burden. Several prominent organizations have developed detailed methodologies to assess the impact of tariffs. Although there are differences between these methods and some unresolved issues, they provide important information to evaluate tariffs.

■ *We would like to acknowledge all of the reviewers at Yale Budget Lab, Penn Wharton Budget Model, CBO, and the Tax Policy Center, as well as the reviewers at JEP.*

References

Amiti, Mary, Stephen J. Redding, and David E. Weinstein. 2019. "The Impact of the 2018 Tariffs on Prices and Welfare." *Journal of Economic Perspectives* 33 (4): 187–210.

Atkin, David. 2024. "Trade and Price-Index Inequality." *Oxford Open Economics* 3 (S1): i1069–75.

Boehm, Christoph E., Andrei A. Levchenko, and Nitya Pandalai-Nayar. 2023. "The Long and Short (Run) of Trade Elasticities." *American Economic Review* 113 (4): 861–905.

Boller, Lysle, Kody Carmody, Jon Huntley, Felix Reichling, and Kent Smetters. 2025. "The Economic Effects of President Trump's Tariffs." Penn Wharton Budget Model, April 10. https://budgetmodel.wharton.upenn.edu/issues/2025/4/10/economic-effects-of-president-trumps-tariffs.

Carloni, Dorian, and Terry Dinan. 2021. "Distributional Effects of Reducing Carbon Dioxide Emissions with a Carbon Tax." Congressional Budget Office Working Paper 2021-11.

Cavallo, Alberto, Gita Gopinath, Brent Neiman, and Jerry Tang. 2021. "Tariff Pass-Through at the Border and at the Store: Evidence from US Trade Policy." *American Economic Review: Insights* 3 (1): 19–34.

CBO (Congressional Budget Office). 2022. *CBO's Use of the Income and Payroll Tax Offset in Its Budget Projections and Cost Estimates.* CBO.

CBO (Congressional Budget Office). 2024. *The Distribution of Household Income in 2021.* CBO.

CBO (Congressional Budget Office). 2026a. "The Budget and Economic Outlook: 2026 to 2036." CBO working paper No. 61882.

CBO (Congressional Budget Office). 2026b. *CBO's Conventional Tariff Analysis Model.* CBO.

Fajgelbaum, Pablo D., Pinelopi K. Goldberg, Patrick J. Kennedy, and Amit K. Khandelwal. 2020. "The Return to Protectionism." *Quarterly Journal of Economics* 135 (1): 1–55.

Fajgelbaum, Pablo D., and Amit K. Khandelwal. 2016. "Measuring the Unequal Gains from Trade." *Quarterly Journal of Economics* 131 (3): 1113–80. https://doi.org/10.1093/qje/qjw013.

Feenstra, Robert C., Philip Luck, Maurice Obstfeld, Katheryn N. Russ. 2018. "In Search of the Armington Elasticity." *Review of Economics and Statistics* 100 (1): 135–150. https://doi.org/10.1162/REST_a_00696.

Gale, William G., and Kyle Pomerleau. 2023a. "Deconstructing the Fair Tax." *Tax Notes Federal* 178 (13): 2169–94.

Gale, William, and Kyle Pomerleau. 2023b. "Consumption Taxes: The Good, the Bad, and the Unworkable." *Journal of Policy Analysis and Management* 42 (3): 846–52.

Ganapati, Sharat, and Colin J. Hottman. 2026. "Did Foreigners Pay America's Tariffs? Quantity Discounts, Scale Economies and Incomplete Pass-Through." NBER Working Paper 34901.

Ghodsi, Mahdi, Julia Grübler, Robert Stehrer. 2016. "Import Demand Elasticities Revisited." Vienna Institute for International Economic Studies Working Paper 132.

Head, Keith and Thierry Mayer. 2014. "Gravity Equations: Workhorse, Toolkit, and Cookbook." In *Handbook of International Economics*, Vol. 4, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, 131–95. North-Holland.

IRS. 2024. *Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2014–2016.* US Department of the Treasury.

Joint Committee on Taxation. 2011. *The Income and Payroll Tax Offset to Changes in Excise Tax Revenues.* JCX-59-11.

Joint Committee on Taxation. 2025a. *Income and Payroll Tax Offset to Changes in Excise Tax Revenues for 2025–2035.* JCX-10-25.

Joint Committee on Taxation. 2025b. "The Joint Committee on Taxation Revenue Estimating Process." Unpublished.

Joulfaian, David, and James Mackie. 1992. "Sales Taxes, Investment, and the Tax Reform Act of 1986." *National Tax Journal* 45 (1): 89–105.

Kalemli-Özcan, Şebnem, Can Soylu, and Muhammed A. Yildirim. 2026. "Global Trade, Tariff Uncertainty and the U.S. Dollar." NBER Working Paper 34728.

Kitchen, John. 2017. "A Technical Note on Revenue Estimation for an Add-On VAT: The Fixed GDP Convention, Relative Prices, Income Offsets, and Monetary Policy." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.3047455.

Li, Huaqun, Garrett Watson, and Erica York. 2025. "Overview of the Tax Foundation's General Equilibrium Model." March 5. https://taxfoundation.org/research/all/federal/general-equilibrium-model/.

Nunns, James R. 2012. "How TPC Distributes the Corporate Income Tax." Tax Policy Center. https://www.urban.org/research/publication/how-tpc-distributes-corporate-income-tax.

Penn Wharton Budget Model. 2026. "Tariff Simulator: Revenue and Prices." University of Pennsylvania. https://budgetmodel.wharton.upenn.edu/issues/2025/2/26/tariff-revenue-simulator (accessed November 2025).

Pomerleau, Kyle, and Erica York. 2025. "Understanding the Effects of Tariffs." AEI Economic Perspectives, April 23. https://www.aei.org/research-products/report/understanding-the-effects-of-tariffs/.

Sutherland, Michael D., and Karen M Sutter. 2025. "China's E-Commerce Exports and U.S. *De Minimus* Policies." Congressional Research Service: In Focus, February 3. https://www.congress.gov/crs-product/IF12891.

Swagel, Phillip L. 2024. "Effects of Illustrative Policies That Would Increase Tariffs." December 18. https://www.cbo.gov/system/files/2024-12/61112-Tariffs.pdf.

Swagel, Phillip L. 2025a. "Budgetary and Economic Effects of Increases in Tariffs Implemented between January 6 and May 13, 2025." June 4. https://www.cbo.gov/system/files/2025-06/61389-Tariff-Effects.pdf.

Swagel, Phillip. 2025b. "An Update about CBO's Projections of the Budgetary Effect of Tariffs." August 22. https://www.cbo.gov/publication/61697.

Swagel, Phillip L., and Thomas Barthold. 2025. "How CBO and Joint Committee Staff Prepare Dynamic Analyses." April 29. https://www.cbo.gov/system/files/2025-04/61346-Dynamic-Scoring.pdf.

Tax Policy Center. 2022. "Brief Description of the Tax Model." March 9. https://taxpolicycenter.org/resources/brief-description-tax-model.

Tax Policy Center. 2025a. "TPC Tariff Tracker." August 12. https://taxpolicycenter.org/features/tracking-trump-tariffs.

Tax Policy Center. 2025b. "T25-0376—Trump Administration Tariffs Announced From January 20, 2025 Through December 4, 2025, Distribution by ECI Percentile, 2026." https://taxpolicycenter.org/model-estimates/T25-0376.

Tax Policy Center. 2026. "T26-0003—Trump Administration Tariffs Announced From January 20, 2025 Through January 8, 2026 With IEEPA Tariffs Overturned, Impact on Tax Revenue, Fiscal Years 2026–2035." https://taxpolicycenter.org/model-estimates/T26-0003.

Toder, Eric, Jim Nunns, and Joseph Rosenberg. 2011. *Methodology for Distributing a VAT.* Tax Policy Center.

Toder, Eric, Jim Nunns, and Joseph Rosenberg. 2012. *Using a VAT to Reform the Income Tax.* Tax Policy Center.

US Census Bureau. 2025. "Guide to the U.S. International Trade Statistical Program." https://www.census.gov/foreign-trade/guide/sec2.html (accessed November 21, 2025).

US International Trade Commission. 2023. *Economic Impact of Section 232 and 301 Tariffs on U.S. Industries*. US International Trade Commission.

Viard, Alan D. 2010. "Sales Taxation of Business Purchases: A Tax Policy Distortion." *State Tax Notes* 56 (12): 967–73.

White House. 2025. "Addressing Certain Tariffs on Imported Articles." Presidential Action, Executive Order, April 29. https://www.whitehouse.gov/presidential-actions/2025/04/addressing-certain-tariffs-on-imported-articles/.

Yale Budget Lab. 2025. "Combined Distributional Effects of the One Big Beautiful Bill Act and of Tariffs." June 12, updated June 16. https://budgetlab.yale.edu/research/combined-distributional-effects-one-big-beautiful-bill-act-and-tariffs.

Yale Budget Lab. 2026. "State of U.S. Tariffs: January 19, 2026." January 19. https://budgetlab.yale.edu/research/state-us-tariffs-january-19-2026.

York, Erica, and Alex Durante. 2026. "Tracking the Economic Impact of the Trump Tariffs and Trade War." Tax Foundation, May 5. https://taxfoundation.org/research/all/federal/trump-tariffs-trade-war/.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 193–212

Correct (and Incorrect) Inference with a Single Instrumental Variable: Practical Takeaways from the Weak Instruments Literature

David S. Lee and Jack Porter

While most economists have some familiarity with "weak instruments," the underlying issues—what a weak instrument is, why it causes problems, what might be done about these problems, and why any particular solution might work—are likely less well appreciated. Moreover, consumers of econometric methods who seek to explore the weak instruments research literature are likely to find it sprawling, specialized, and technically dense.

This article offers a rudimentary introduction to the "weak instruments" problem, including some practical do's and don'ts implied by the findings of the weak instruments research literature. Our starting point is to explain why the standard approaches to statistical inference do not give correct inferences for linear instrumental variables. This leads to our discussion of existing solutions to the problem. Throughout the paper, we will focus on the simple "just-identified" instrumental variables model with one regressor and one instrument. We will also refer specifically to the confidence level of 95 percent and the hypothesis testing significance level of 5 percent, because these are default standards in applied work, but the discussion also applies to other confidence or significance levels.

To clarify what we mean when we say the standard approaches for the instrumental variable model "do not give correct inferences," we recall an issue that

■ *David S. Lee is the Chemical Bank Chairman's Professor of Economics and Public Affairs, Princeton University, Princeton, New Jersey. He is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Jack Porter is Richard E. Stockwell Distinguished Chair and Professor of Economics, University of Wisconsin, Madison, Wisconsin. Their email addresses are davidlee@princeton.edu and jrporter@ssc.wisc.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20251464.

is more familiar to researchers. In standard regression models, the errors in the model are either homoskedastic—sometimes referred to as "equal scatter"—or non-homoskedastic (heteroskedastic and/or correlated across observations), and there are standard error formulas that are appropriate for each case. Importantly, the homoskedastic standard error formulas provide incorrect inferences if non-homoskedastic errors are present, but the so-called "robust standard errors" ("White," "Newey-West," or "clustered") lead to correct inferences whether or not the errors are homoskedastic. Because researchers do not know the true nature of the errors, it has become the safe and standard practice to compute and report robust standard errors, the ones that work in either case.

In the same way that homoskedasticity-only standard errors are inappropriate and incorrect when the true errors are not homoskedastic, the standard two-stage least squares estimator and usual standard errors are inappropriate and incorrect when "instruments are weak." And as we discuss below, just as it is not possible to know with certainty whether or not errors are homoskedastic in the standard regression model, it is not possible to know with certainty whether or not instruments are problematically weak. The good news is that the econometric theory literature has proposed solutions—so-called "robust-to-weak-instruments" methods—to allow for correct inference. The bad news is these practical solutions are to some extent buried in a technically dense theoretical literature, creating a gap between common practice and the theoretical literature's most basic and long-standing recommendations.

We start with a quick review of linear instrumental variables and weak instruments, using the estimation of the returns to education as a running example. We then explain in the context of an instrumental variable model with a single excluded instrument, (1) why the usual "t-statistic" and the interval that uses ± 1.96 times the two-stage least squares standard error leads to incorrect inferences, (2) what procedures constitute fixes for this problem, and (3) what are some inferential pitfalls to avoid when using the first-stage F statistic in this context.

A Running Example: Returns to Education

The linear instrumental variable model is an important part of the modern empirical economics toolkit. Researchers commonly use it when omitted variable bias, endogeneity, or measurement error threatens the validity of standard regression models. As a running example, imagine a researcher who is interested in estimating the causal effect of education in a linear regression with hourly earnings as the dependent variable and years of education as the key explanatory variable (along with various control variables). An obvious concern is that years of education may be dependent on other potentially unobservable variables, like persistence or parenting or family connections, but might also contribute to earnings capacity, leading to an ambiguous interpretation of the standard linear regression coefficient, which could reflect the true impact of education, or alternatively some of these other unobserved factors.

The instrumental variable approach offers one possible way to address this problem. It is typically described in three equations. The first equation, often called the main equation, describes how changes in the X variable affect the Y variable: for example, how do years of education affect hourly earnings? In this equation, the key parameter of interest is β.

$$Y = \alpha + X\beta + u$$

$$X = \gamma + Z\pi + v$$

$$Y = \underbrace{\alpha + \gamma\beta}_{\delta_0} + Z\underbrace{\pi\beta}_{\delta_1} + \underbrace{u + v\beta}_{\varepsilon}$$

The key causal inference problem is captured by the possibility that the error term u (unobserved factors such as persistence or family connections) might be correlated with X.

The second equation, called the first-stage equation, describes how a plausibly exogenous variable Z causes some of the variation in the endogenous variable X. In order for the instrumental variable strategy to work, it must also be assumed that Z—the instrument—is uncorrelated with u. The well-known analysis of Card (1995) uses "college proximity" as this Z variable for estimating the effect of years of education on hourly earnings, building on the intuition that those who live farther from a college are less likely to attend, for reasons uncorrelated with underlying unobserved factors that otherwise ultimately impact hourly earnings.

A third important equation is a combination of the first two: the "reduced form" equation. Algebraically, it substitutes the expression for X in the first-stage equation into the main equation. It thus describes the relationship between the dependent variable Y and the exogenous instrumental variable Z, as mediated through the years of education variable. In this equation, the terms can be grouped into a composite constant term δ_0, the term involving Z, including the slope coefficient δ_1, plus an error term. The δ_1 coefficient, in particular, will be the product of the coefficients from the main and first stage equations. In the returns to school example, δ_1 essentially represents the impact of college proximity on earnings through its impact on years of schooling. To account for additional control variables in the analysis, one includes the same set of control variables in all three equations.[1]

[1] Note that the parameter β can be viewed as the ratio of two population covariances, $\frac{C(Y,Z)}{C(X,Z)}$, which is the parameter of interest in many other contexts: the slope coefficient in a simple linear regression model, where X is correlated with the error, and Z is not; the so-called "Local average treatment effect" in the context of heterogeneous treatment effects, when X is a binary treatment variable and Z is a binary instrument; or the regression coefficient of interest when X mismeasures a true regressor of interest X^*, and Z is an alternative measure of X^* with independent measurement errors. In the "fuzzy regression discontinuity" or "fuzzy regression kink design," the parameter of interest is an analogous ratio of two regression coefficients.

Table 1
Replication of Table 3, Panel A, from Card (1995)

	Reduced form models		
	Education (2)	Earnings (4)	Structural models of earnings (6)
Live near college in 1966	0.322	0.045	
	(0.083)	(0.018)	
Education			0.140
			(0.055)
tF standard error			(0.080)
AR confidence interval			[0.037, 0.291]
$\hat{r}$			−0.319
First-stage *F*			15.086
Family background variables	Yes	Yes	Yes

Source: Card (1995).
Note: $N = 3{,}010$. Information in black text is taken from panel A of Table 3 of Card (1995). Details of included covariates are found in Card (1995). The information in red text is new, relative to the original text. "*tF* Standard Error" is the adjusted standard error using the adjustments in Lee, McCrary, Moreira, and Porter (2022) for a 5/95 percent test/confidence interval. "*AR* confidence interval" is the 95 percent confidence interval using Anderson and Rubin (1949). $\hat{r}$ is the empirical correlation between the two-stage least squares main equation residual $\hat{u}$ and the first stage residual $\hat{v}$. "First-stage *F*" is the square of the *t*-ratio that tests the null of no effect in the first stage.

Throughout this article, we stipulate that Z is indeed uncorrelated with u, so that it is a valid instrument, even though the error term v (other non-college proximity determinants of education) (and therefore X, too) might be correlated with u. The issue that we focus on is really a statistical inference problem: assuming that the needed model assumptions are correct, how do we make statements about β using sample data?

In his study of the returns to education, Card (1995) uses data from the National Longitudinal Survey of Young Men Cohort (Card n.d.), Y is log hourly earnings, X is years of education, and the instrument Z is the indicator variable for whether the individual lived near a college.

The text in black in Table 1 replicates selected columns from panel A, Table 3, from Card (1995). Column 2 reports the estimate of π from estimating the first-stage regression equation: living near a college is associated with an increase in 0.322 years of education. Column 4 reports the estimate of δ_1 and standard error from the reduced-form equation: living near a college is associated with an approximately 4.5 percent increase in hourly earnings.

Column 6 reports the estimate of β in the main equation using the method of two-stage least squares. It starts by estimating the first-stage equation. It then uses the coefficient from that first-stage equation to create predicted values of X, call it $\hat{X} = \hat{\gamma} + Z\hat{\pi}$—the predicted schooling level based on college proximity. These predicted values can then be used in place of X in the main equation in a

second-stage regression, where the resulting estimated coefficient on $\hat{X}$ is a consistent estimate of the true value of β of 0.140. The two-stage least squares coefficient on education in column 6 is numerically equivalent to the ratio of the coefficient in 4 divided by the coefficient in 2.

By default, popular statistical packages report the standard errors (robust to heteroskedasticity, or clustering). It is implied that these standard errors, along with the two-stage least squares estimates (as shown here in Table 1), can be used to compute *t*-ratios for hypothesis testing or 95 percent confidence intervals using the estimate ± 1.96 times the standard error. But theoretical advances in the econometric literature since this 1995 study have established that these *t*-ratios or confidence intervals are as incorrect as using "homoskedasticity-only" standard errors when heteroskedasticity is present in the standard linear regression model. This problem is at the core of the so-called "weak instruments" theoretical literature.

The Essence of the "Weak Instruments" Inference Problem: A Graphical Illustration

Before describing the valid solutions that have been developed for the problem of weak instruments, it is useful to gain an appreciation as to why the standard methods that use the default output of popular statistical packages—a *t*-ratio or a "± 1.96 times standard error"-type confidence interval—leads to erroneous inferences.

In essence, the reason for this has its roots in the fact that the distribution of the *t*-ratio for the instrumental variable model is actually different from that of the *t*-ratio in a standard regression model. In a standard regression model, when the null hypothesis is true, the *t*-ratio is known (with sufficiently large samples) to closely approximate a standard normal distribution. By contrast, with the instrumental variable model, even with very large samples, the analogous *t*-ratio is *not* closely approximated by the standard normal. Instead, the shape of this non-normal distribution follows that of a nonlinear function of normal random variables and is governed by two parameters.

The first parameter is the expected value of the *F*-statistic. The *F*-statistic itself is computed by taking the estimated coefficient for π, divided by its standard error, and squaring the result: that is, $F \equiv \left(\frac{\hat{\pi}}{\widehat{se}(\hat{\pi})}\right)^2$. This *F* statistic can be thought of, for example, as a sample measure of "how statistically significant" the relationship is between education and proximity to college. $E[F]$, on the other hand is the true and unknown strength of this relationship. It will turn out that the smaller is $E[F]$, the more non-normal the *t*-ratio becomes.

The second parameter is the correlation—which we denote ρ—between the error u in the main equation (non-schooling determinants of earnings), and the error v in the first-stage equation (non-college proximity determinants of schooling). This parameter can be viewed as the "degree of endogeneity." When this correlation is zero, there is no problem estimating the main equation with ordinary regression.

But more generally, the presumption of using instrumental variable methods is that we do not know what ρ might be. It turns out that the larger the magnitude of ρ, the more distorted the *t*-ratio becomes. To be clear, the fact that the "degree of endogeneity" is unknown is precisely why we are using an instrument in the first place. Its magnitude is irrelevant with infinite sample sizes, but becomes a crucial determinant of statistical distributions when sample sizes are finite.

The influence of $E[F]$ and ρ on the distribution of the *t*-ratio can be easily demonstrated from a Monte Carlo simulation. We construct 10,000 independently drawn random samples of size 1,000 following the linear instrumental variable model represented by the main equation and the first-stage equation. We model the simulation after the Monte Carlo example from Angrist and Pischke (2009) (p. 210): from the first-stage regression, the instrumental variable Z and its error term v are drawn so that each has a standard normal distribution, $\beta = 1$; the correlation ρ between the two error terms u and v is 0.2 or 0.8. And the first-stage regression is set for the scenarios $E[F] = 5$ (relatively weak instrument) and $E[F] = 20$ (relatively strong instrument).

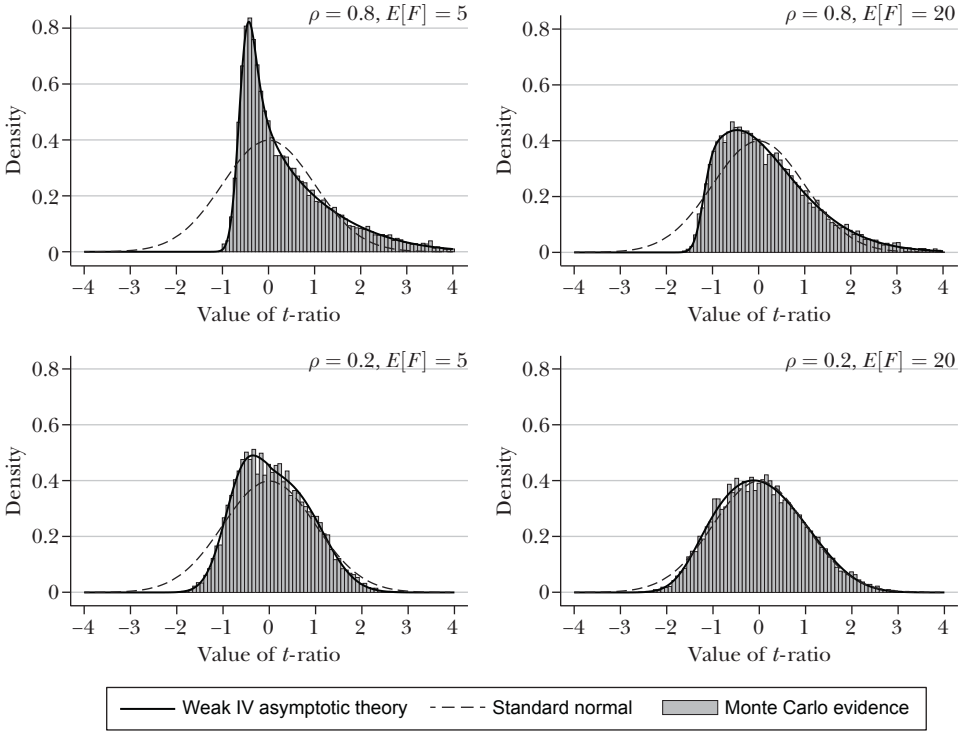
We compute the *t*-ratio using the standard two-stage least squares and standard errors that are reported by default from common statistical packages, and, in Figure 1, plot the distributions of these *t*-ratios across the 10,000 draws, one for each of the four scenarios that cover relatively "low" and "high" endogeneity, and relatively "weak" and "strong" instruments.[2]

In order for the standard methods to give correct inferences, the distributions would need to follow the standard normal distribution, represented by the dashed line in the figure. The basis for using ± 1.96 as critical values for a "5 percent test" is that a normally distributed *t*-ratio will exceed 1.96 in magnitude 5 percent of the time. It is clear from Figure 1 that, in the case of the instrumental variable model, the distributions of the *t*-ratios do not follow a normal distribution, with departures from normality especially more pronounced for smaller $E[F]$ and larger ρ. And herein lies the problem: while some combinations of ρ and $E[F]$ might reject at less than 5 percent of the time, others will lead to a rejection of the null more than the intended 0.05 rate. Indeed, in extreme cases, the rejection rate can reach close to 100 percent! The possibility of misleadingly over-rejecting (or equivalently, the confidence interval "under-covering") is arguably *the* cardinal sin of statistical inference.

Two questions naturally arise in light of these facts. First, are these non-normal distributions an artifact of small sample sizes (here, 1000)? The answer is "no." It is easy to show that the same picture would emerge from an exercise that used sample sizes of 100,000, 1 million, or 10 million. As long as the measure of instrument strength is kept constant—here with an expected value of F of 5 or 20—the resulting histogram will never converge to a normal, and instead its non-normal shape will barely change at all.

[2]For this exercise, we slightly modify the stata simulation code that is included in the supplementary material to Lee, McCrary, Moreira, and Porter (2022), at http://www.princeton.edu/davidlee/wp/SupplementarytF.html.

Figure 1
Distribution of the instrumental variable *t*-ratio



Source: Authors' calculations.
Note: Each panel shows the simulated distribution of the two-stage least squares t-ratio (testing the true β) over 10,000 samples of size 1,000, for one pairing of the endogeneity parameter ρ and the expected first-stage F, $E[F]$. The histogram is the Monte Carlo distribution; the solid line is the Staiger and Stock (1997) weak-instrument asymptotic density; the dashed line is the standard normal.

Second, can we simply definitively rule out the combinations of $E[F]$ or $|\rho|$ that cause any problems? For example, it is tempting (but incorrect) to use the following logic: "My first-stage F is really large. I conclude that I don't have a 'weak instrument,' so the 'weak instrument problem' (of underlying non-normal distributions) doesn't apply to me." The problem with this logic is that the *observed* first-stage F statistic is not the same thing as the *expected value* of F. The observed first-stage F is a specific realization from an underlying random variable, while the true value of $E[F]$ is unknown. Indeed, an F statistic value of 10 is roughly equally likely to be observed when the true $E[F]$ is 5 as when the true $E[F]$ is about 20. As another example, it is extremely unlikely to observe an F of 10 if the true $E[F]$ is 38, a seemingly "strong instrument"; but it is equally unlikely to observe the value of 10 if the instrument Z had no correlation with X ($E[F] = 1$).

Analogously, it would be tempting to use the following logic using the observed correlation between the residuals $\hat{u}$ and $\hat{v}$, which we denote $\hat{r}$: "I observe a small

(in magnitude) $\hat{r}$, so the ρ is small enough for me to ignore the problem of weak instruments." As was the case with the F statistic, this is erroneous reasoning: $\hat{r}$ is an *estimate,* and therefore we are unable to use it to make definitive conclusions like this one about the true and unknown ρ.

Intuitively, while statistics like F (a sample estimate of the association between schooling and college proximity) and $\hat{r}$ (a sample estimate of endogeneity) are important statistics (as we will discuss below) that help us with the problem, they are no more a valid substitute for unknown parameters like $E[F]$ or ρ than the two-stage least squares estimate is a substitute for the true β. As with any statistical inference problem, the best we can do is to use data to estimate and make statistical inferences about parameters like β, ρ, or $E[F]$.

The non-normal distributions associated with weak instruments in Figure 1 could give an impression that "all bets are off" in this instrumental variable problem. But the opposite is true. Staiger and Stock (1997) developed a way to accurately predict the shape of these non-normal distributions. Indeed, the solid line in Figure 1 depicts the density implied by their theoretical framework and derivations, and it very closely matches what happens in the Monte Carlo simulations. And furthermore, Moreira (2003, 2009) developed a framework for constructing tests of statistical significance that would be correct—or in the parlance of the "weak instruments" literature—tests that would be "robust to weak instruments." The phrase "robust to weak instruments" has a meaning that is directly comparable to the phrase "robust to heteroskedastic errors." It is now standard practice to employ "heteroskedasticity-consistent" standard errors because researchers wish not to take a stand on whether or not the errors in a regression are homoskedastic. Similarly, a statistical test or confidence interval procedure that is "robust to weak instruments" is a procedure that does not require you to take a stance on the unknown true magnitudes of ρ or $E[F]$.

Methods for Calculating Confidence Intervals with Weak Instruments

In this section we highlight some practical takeaways from the weak instruments literature. We begin by highlighting the key statistics that capture all the relevant information about a particular problem and are sufficient for producing valid inference. All of the valid inference procedures implicitly use the information completely captured by these statistics. We then describe some of these valid inference methods, and then move to discussing some potential misuses of the information and misinterpretations to avoid.

Just as a solution to the problems caused by heteroskedasticity in the linear regression model is to use clustered or heteroskedasticity-robust standard errors, the most straightforward solution to the inferential problems caused by the possibility of weak instruments is to use an inference method that works irrespective of the true strength of the instrument or degree of endogeneity, a so-called

"robust-to-weak-instrument" method. As we will discuss, there is a whole *class* of test procedures and confidence intervals that produce correct inference without taking any stand on the true value of $E[F]$ or ρ. These robust-to-weak-instruments inference methods produce usable and potentially informative confidence intervals even if the first stage F is as low as $1.96^2 = 3.84$, which is the border of statistical significance of the first stage.

Compute and Preserve Five Necessary Statistics from the Three Regressions

As a preliminary step, our first recommendation is to compute and preserve five statistics from the main equation, the first-stage equation, and the reduced form equation. These five statistics contain essentially all the relevant information used by any of the procedures for robust estimation of statistical confidence intervals in the context of weak instruments literature (including the ones described below).

The first two statistics come from estimating the main equation using two-stage least squares, with Z as the instrumental variable: (1) the two-stage least squares coefficient $\hat{\beta}_{IV}$, and (2) its associated standard error. The next two statistics come from estimating the first-stage regression: (3) the coefficient $\hat{\pi}$ and (4) its standard error. The fifth necessary coefficient comes from estimating the reduced form equation: (5) the standard error for the reduced-form coefficient (which is itself mechanically equal to the product of $\hat{\beta}_{IV}$ and $\hat{\pi}$).

The Card (1995) study reports these five statistics transparently, allowing one to verify that the division of the reduced form coefficient by the first-stage coefficient equals the two-stage least squares coefficient, and uses a consistent standard error formulation for all three regressions. In more recent studies, the reporting of all five of these statistics occurs to some extent, but not consistently so. For example, Lee, McCrary, Moreira, and Porter (2022) find that in their sample of 61 papers published in the *American Economic Review* that only about one-quarter of the instrumental variable specifications contain all five statistics.

Making available these five statistics allows researchers to implement existing robust methods, as well as new procedures that may be developed in the future. Moreover, these five sufficient statistics can be made public without any fear of compromising the privacy of any observation in the dataset.

From these five statistics, one can compute the first-stage estimate $F \equiv \left(\frac{\hat{\pi}}{\widehat{se}(\hat{\pi})}\right)^2$ for $E[F]$, as well as the estimate $\hat{r}$ for ρ. The formula for $\hat{r}$ is more involved.[3]

In Table 1, we have added the statistics F and $\hat{r}$ for the Card (1995) example in red text. In column 6, the first-stage F is 15.086. This tells us that the relationship between schooling and college proximity is statistically different from zero at conventional levels of significance (even if we cannot definitively rule out no relationship). Meanwhile the empirical correlation between the two-stage least

[3]Specifically, $\hat{r} = \frac{|\hat{\pi}|\hat{\beta}_{IV}}{2\widehat{se}(\hat{\pi})\widehat{se}(\hat{\beta}_{IV})}\left(\frac{\widehat{se}(\widehat{\pi\beta})^2}{\hat{\pi}^2\hat{\beta}_{IV}^2} - \frac{\widehat{se}(\hat{\pi})^2}{\hat{\pi}^2} - \frac{\widehat{se}(\hat{\beta}_{IV})^2}{\hat{\beta}_{IV}^2}\right)$.

squares residual $\hat{u}$ and the first-stage residual $\hat{v}$ is -0.319, an estimate of the true and unknown degree of endogeneity, ρ.[4]

The Standard Robust-to-Weak-Instrument Inference Method: Anderson and Rubin (1949)

The method of Anderson and Rubin (1949) (*AR*) is the standard robust-to-weak-instrument inference method for the single instrumental variable case. Although the Anderson-Rubin test had existed for over 75 years, Staiger and Stock (1997) first showed that it correctly dealt with the weak instrument problem. The Anderson-Rubin approach has more recently been a focal point of overviews and survey articles intended for a broader audience (for example, Andrews and Stock 2018; Andrews, Stock, and Sun 2019; Keane and Neal 2023, 2024), but has been slow to show up in introductory or more accessible econometrics texts (with a few exceptions, like Stock and Watson 2019; Cameron and Trivedi 2022).

A hypothesis test at the 5 percent level using the Anderson-Rubin approach involves computing the quantity $AR \equiv \left(\frac{\widehat{\pi\beta} - \hat{\pi}\beta_0}{\widehat{se}\left(\widehat{\pi\beta} - \hat{\pi}\beta_0\right)}\right)^2$ and rejecting the hypothesis that $\beta = \beta_0$ if AR exceeds $1.96^2 = 3.84$.[5]

Applied researchers are typically interested in the broader question of the *range* of values of β that would or would not be statistically rejected by the data—that is, a confidence interval. Analogous to the standard ± 1.96 confidence interval for basic regressions, it is straightforward to produce the Anderson-Rubin confidence interval. The Anderson-Rubin 95 percent confidence interval bounds will be: $\left[\hat{\beta}_{IV} + k_l(\hat{r}, F) \cdot \widehat{se}\left(\hat{\beta}_{IV}\right), \hat{\beta}_{IV} + k_u(\hat{r}, F) \cdot \widehat{se}\left(\hat{\beta}_{IV}\right)\right]$, where k_l and k_u are known functions of F and $\hat{r}$. As reported in Section C.1 of the online appendix to Lee et al. (2022), k_l and k_u will be:

$$k_l = \frac{-\frac{1.96}{\sqrt{F}}\hat{r} - \sqrt{1 - \frac{1.96^2\left(1-\hat{r}^2\right)}{F}}}{1 - \frac{1.96^2}{F}} \cdot 1.96, \quad k_u = \frac{-\frac{1.96}{\sqrt{F}}\hat{r} + \sqrt{1 - \frac{1.96^2\left(1-\hat{r}^2\right)}{F}}}{1 - \frac{1.96^2}{F}} \cdot 1.96.$$

Note how as F becomes very large, k_l, k_u become close to the standard ± 1.96 values, but for moderate or smaller values of F, they can depart quite a bit from the ± 1.96 values. By plugging in $\hat{r} = -0.319$ and $F = 15.086$ into the above formulas, one obtains the standard error multiples $k_l = -1.89$ and $k_u = 2.73$. Using the standard error of 0.055, the Anderson-Rubin confidence interval is

[4] A spreadsheet illustrating how these calculations are performed can be found at www.princeton.edu/~davidlee/wp/ivcalc.html.

[5] The denominator in this quantity is a bit more involved, but computable as $\widehat{se}\left(\widehat{\pi\beta} - \hat{\pi}\beta_0\right) = \left[\widehat{se}\left(\widehat{\pi\beta}\right)^2 + \beta_0^2\widehat{se}\left(\hat{\pi}\right)^2 - \frac{\beta_0}{\hat{\beta}_{IV}}\left(\widehat{se}\left(\widehat{\pi\beta}\right)^2 + \hat{\beta}_{IV}^2\widehat{se}\left(\hat{\pi}\right)^2 - \hat{\pi}^2\widehat{se}\left(\hat{\beta}_{IV}\right)^2\right)\right]^{1/2}$.

computed for the Card (1995) study and shown in red text in Table 1 as the interval (column 6) $[0.037, 0.291]$.[6]

Practitioners should be aware of two features of the Anderson-Rubin confidence interval. First, it will generally not be centered around the two-stage least squares estimate (as is evident in Table 1), in which the point estimate is 0.14, which is closer to the endpoint 0.037 than it is to the endpoint 0.291. This may seem unfamiliar or unusual, but on the other hand, there is probably no reason to have expected it to be centered, given that robust-to-weak-instruments procedures need to account for asymmetries in distributions in the problem (as shown earlier in Figure 1). Aiming for full transparency suggests reporting both $\hat{\beta}_{IV}$ and the Anderson-Rubin 95 percent confidence interval endpoints.

Another way to make sense of the asymmetry of the *AR* confidence interval is to note that these asymmetric inflation factors imply a numerically equivalent way to conduct the *AR* test: use the standard two-stage least squares *t*-ratio, but instead of using the critical values -1.96 and 1.96, use the critical values $-k_u$ (a negative value, here -2.73) and $-k_l$ (a positive value, here 1.89).

Note that in Figure 1, ρ is positive, and the *t*-ratio is skewed to the right; if ρ were negative, the resulting distribution would be a mirror image, skewed to the left. We would expect correct critical values to similarly be skewed to the left when the empirical counterpart $\hat{r}$ is also negative. And this is indeed what we see in the Card example, with critical values -2.73 and 1.89. In the Card example, the test of the null hypothesis that $\beta = 0$ would imply a *t*-ratio of $\frac{0.14 - 0}{0.055} = 2.55$, a value that exceeds 1.89 and therefore is statistically significant. This, of course, is the same thing as observing that 0 is outside the reported Anderson-Rubin interval $[0.037, 0.291]$.

A second feature worth noting is that the Anderson-Rubin 95 percent confidence set is a bounded interval if and only if $F > 1.96^2$. In particular, if $F \leq 1.96^2$, then the *AR* confidence set would be unbounded (and the k_l and k_u above could not be used to form a bounded interval). This may initially seem strange, but as has been pointed out in other surveys (for example, Andrews, Stock, and Sun 2019; Keane and Neal 2023, 2024), this should be considered a feature, not a bug, and is actually quite intuitive. As Andrews, Stock, and Sun (2019) put it, "[I]nfinite-length confidence sets arise exactly in those cases where the data do not allow us to conclude that β is identified at all"—when we cannot even statistically reject that $\pi = 0$, the case when the instrument is completely irrelevant. This intuitive feature aligns with one of the earliest results in the literature on weak instrumental variables, that the possibility of an unbounded confidence set has been shown to be a *necessity* to having a valid confidence set procedure, in the context of the instrumental

[6]As of STATA 19.0, the 95 percent confidence interval can be computed by invoking "ivregress 2sls y w [x=z]," followed by "estat weakrobust, ci," where *y* is the dependent variable, *w* is a control variable, *x* is the endogenous regressor of interest, and *z* is the instrument.

variable model (Dufour 1997).[7] In other words, one way to immediately tell whether a confidence interval procedure is invalid is if it always produces a bounded interval (like the usual ± 1.96 interval formulation, as well as typical Wald bootstrap methods for computing confidence intervals).

In summary, the correct use of the Anderson-Rubin method for a 95 percent level of confidence is as follows: the researcher examines the first stage F, if it is less than $1.96^2 = 3.84$, then the conclusion is that neither infinity nor negative infinity can be statistically ruled out as a possible value for β: the instrument is uninformative in that it does not produce a bounded confidence interval. If $F > 3.84$, then the 95 percent confidence interval can be constructed as described above.

In the case of the Card (1995) study, focusing on column 6 of Table 1, the ratio of the length of the *AR* interval to the usual 1.96 interval is $\frac{0.291 - 0.037}{2 \cdot 1.96 \cdot 0.055} \approx 1.18$, suggesting that the correction via the Anderson-Rubin interval leads to about an 18 percent wider confidence interval than the usual 1.96 interval.

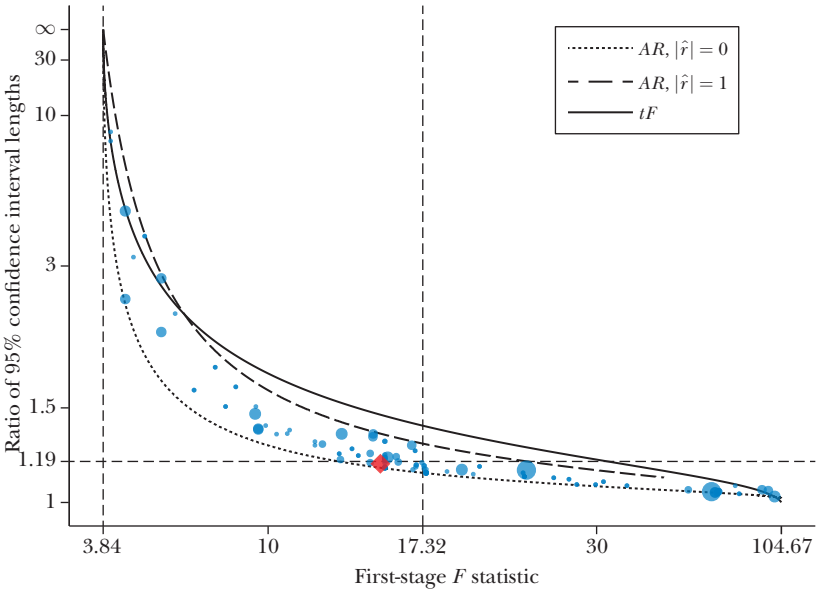
It is natural to wonder whether this degree of overstating of precision is an outlier or might be typical in applied research. To answer this question, we use the dataset (Lee and Porter 2026) of single instrumental variable specifications collected by Lee et al. (2023). The studies are drawn from articles published in the *American Economic Review*, *Quarterly Journal of Economics*, *Journal of Political Economy*, *Econometrica*, and *Review of Economic Studies*. These studies reported the five statistics needed for computing the Anderson-Rubin confidence intervals.

In Figure 2, using this sample of studies, we plot the first-stage F statistic for each study on the horizontal axis, and the ratio of the length of the Anderson-Rubin confidence interval to that of the usual 1.96 interval against the first-stage F statistic.[8] For reference, we also show with the long dash and dotted line, the full theoretical range of the ratio confidence interval lengths, using the extreme possible values of the correlation between the error term $|\hat{r}|$—that is, either zero or one. Notice that when a study's F statistic is very large, the ratio is essentially equal to 1. The weighted median F statistic for these studies is 17.32: the F statistic from the Card (1995) study is close to the median in this sample of studies. There is quite a bit of range of F statistic values across the studies, and as intuition would suggest, as F decreases (less confidence about whether the first-stage coefficient π is nonzero), the relative length of the Anderson-Rubin interval increases, as does the extent to which the usual 1.96 interval overstates precision. This degree of overstatement explodes to infinity as the F statistic approaches 3.84. For F values around 10—a commonly

[7] For example, when $F < 3.84$ the *AR* confidence set can be $(-\infty, \infty)$, or it can be the union of two disjoint sets, $(-\infty, a] \cup [b, \infty)$, where a and b are the resulting endpoints by using the formula for the interval endpoints above.

[8] If all studies had the same true value of $E[F]$, then the distribution of F statistics would approximately follow a non-central chi-squared distribution with one degree of freedom, with the same non-centrality parameter. To capture differences in the non-centrality parameter, we fit the F statistics to a Weibull distribution and use that cumulative distribution function to transform the F for the plot. This is approximately using the rank of the F statistic as the x-axis.

Figure 2
Anderson-Rubin Interval Lengths and *tF* Interval Lengths for Instrumental Variable Specifications



Source: Lee and Porter (2026).
Note: Vertical axis measures the ratio of the length of the *AR* (or *tF*) 95-percent confidence interval to the length of the usual two-stage least squares confidence interval, $2 \times 1.96 \times \hat{\text{se}}(\hat{\beta}_{\text{IV}})$. Blue circles are the 89 specifications used for Figure 6 of Lee et al. (2023), where the areas of the circles are proportional to the inverse of the number of specifications within a study. The red diamond represents the additional specification from the Card (1995) specifications from Table 1. The scale for the vertical axis is a logistic transformation of ln(ratio) (that is, ratio/(1+ratio)). The scale for the horizontal axis uses a Weibull cumulative distribution function transformation with scale parameter 21.65 and shape parameter 1.11. The ratio of confidence interval lengths of $\frac{1.96}{1.645} = 1.19$ is provided as a benchmark.

cited (but misused, as we discuss below) reference value—this correction ratio can range from 1.27 to 1.62, depending on the value of $\hat{r}$.

Finally, to provide some perspective on the degree of distortion in statistical conclusions, we note that the ratio of 1.19 would be the ratio $\left(\frac{1.96}{1.645}\right)$ of the length of a 95 percent confidence interval to that of a 90 percent confidence interval. In other words, for roughly half of the specifications in the sample (those with first-stage *F* statistics less than the median of 17.32), use of the conventional ±1.96 confidence interval instead of the valid Anderson-Rubin interval is akin to overstating the precision *more* than using a 90 percent confidence interval but mislabeling it as a 95 percent confidence interval in more standard regression problems.

The *tF* Inference Method: Lee, McCrary, Moreira, and Porter (2022)

The Anderson-Rubin (1949) method is not the only robust-to-weak-instrument inference method. Taking a closer look at how the first-stage *F* statistic can help

correct inference, Lee, McCrary, Moreira, and Porter (2022) derived a function of F, $\sqrt{c(F)}$, to replace 1.96 such that the usual two-stage least squares t-ratio could be used for inference. For this method—called *tF*, referring to a procedure that only uses the t-ratio and the first-stage F statistic—the 95 percent confidence interval is computed easily as $\hat{\beta}_{IV} \pm \sqrt{c(F)}\,\widehat{se}\left(\hat{\beta}_{IV}\right)$. The function $\sqrt{c(F)}$ does not have an explicit functional form, but a table of critical values is reported in Lee, McCrary, Moreira, and Porter (2022): for the F value of interest, one simply looks up the value $\sqrt{c(F)}$. Alternatively, the set of appropriate values are implemented in a STATA.ado file available for download.[9]

In terms of a minimal standard of delivering correct inferential statements—irrespective of the unknown values ρ and $E[F]$—both Anderson-Rubin and *tF* equally meet that standard. There are, however, differences between the two methods that have some practical implications.

First, because it does not utilize the information contained in the five statistics referenced above, we expect *tF* to yield wider confidence intervals. This expectation is visible in Figure 2, which plots the ratio of *tF* confidence interval length to the usual 1.96 interval length. We see that for a broad range of F values the Anderson-Rubin confidence intervals are shorter. As shown in the figure, depending on the magnitude of $\hat{r}$, the Anderson-Rubin confidence intervals can be longer at smaller values of F and larger values of F (beyond $F = 104.67$).[10] Generally, the use of all the relevant information in the problem can often provide a clear advantage of *AR*.

On the other hand, many past studies do not report the five statistics needed to recover the Anderson-Rubin interval, but do report the first-stage F statistic, leaving the question of how to adjust findings of these past studies to obtain valid inference without obtaining access to the original microdata. *tF* is ideally suited for this situation, as it only requires the use of the first-stage F statistic. Additionally, when one would like "apples-to-apples" comparisons of degrees of precision between two studies, one of which only reports the first-stage F, *tF* can be used to make that comparison.

Second, in the *tF* approach, the scaling factor for the standard error is actually the usual ± 1.96 when F exceeds 104.67. That is, *tF* can be effectively combined with the usual confidence interval in the following way: use the usual ± 1.96 interval if $F > 104.67$, and the $\pm\sqrt{c(F)}$ interval otherwise. By contrast, as shown in Lee, McCrary, Moreira, and Porter (2022), it is *not* valid to combine the Anderson-Rubin interval and the usual 1.96 interval in a similar way. The Anderson-Rubin approach works just fine on its own, but one cannot standardly use it when F is small and then shift to the usual ± 1.96 method when F is large, as tempting as that might be.

[9] See http://www.princeton.edu/davidlee/wp/SupplementarytF.html.

[10] It actually turns out that the repeated sample distribution of the length of *AR* confidence intervals has a thick upper tail, so much so that the average length is infinite; meanwhile, the average length for *tF* confidence intervals is finite (Lee, McCrary, Moreira, and Porter 2022). The average here is conditional on $F > 1.96^2$, when both *AR* and *tF* intervals are bounded. So from an average length of confidence interval perspective, *tF* dominates *AR*.

Finally, *tF* may be more easily explained to nonspecialists as a simple correction to the usual "introductory textbook" methods. Put simply, using ± 1.96 to scale the standard error of the $\hat{\beta}_{IV}$ variable ignores the statistical uncertainty about $\hat{\pi}$, while using $\pm\sqrt{c(F)}$ does not. For an audience outside academic circles that is expecting a single "standard error" in reporting, the explicit formula for the Anderson-Rubin confidence interval endpoints may be hard to digest and may seem somewhat opaque. By contrast, *tF* allows reporting results in the familiar format of a point estimate $\hat{\beta}_{IV}$ with the corresponding standard error in parentheses, with the latter replaced by a corrected "5 percent *tF* standard error." We have illustrated this way of reporting *tF* standard errors in Table 1 in the red text for the Card (1995) study.

Our own view is that any arguments for Anderson-Rubin or *tF* methods based on precision, or reporting convenience, should not lose sight of the fact that both are fully robust to weak instruments.

Other Robust-to-Weak-Instrument Methods

One key insight from the weak instruments literature is that there are not just one or two solutions to the inference problem; instead, there exists a whole *class* of robust-to-weak-instrument solutions that deliver correct statistical significance statements and confidence intervals (Moreira 2003, 2009). Here, we briefly call attention to two other robust inference methods that have been proposed for the single instrument case.

In the "Conditional Wald" (*CW*) method of Moreira (2003), the concept of the test is quite straightforward. One computes the two-stage least squares *t*-ratio as usual and rejects the null hypothesis when t^2 (the Wald statistic) exceeds a critical value c_{CW}, which itself is a function of the five key quantities mentioned above and β_0, the null hypothesis value.

Two practical obstacles have limited the use of the Conditional Wald approach in practice. First, c_{CW} does not have an explicit functional form, and so must be determined by some numerical or Monte Carlo method; indeed, these critical values must eventually be presented in a multidimensional table with choices to be made about the degree of resolution. Second, inverting the test to produce confidence intervals is not a trivial task. One can essentially manually test hypothesized values of β on a discretized grid of values, but such an implementation of *CW* carries the caveat that there are approximation errors that are not easily formally quantified.

Both numerical complications also arise with another method, recently proposed by Lee et al. (2023): the *VtF* procedure, where the "V" differentiates it from the *tF* procedure in the sense that it uses all five of the statistics that Anderson-Rubin and Conditional Wald use. It should be considered a close cousin to the Conditional Wald approach in that the test rejects the null when t^2 exceeds a critical value c_{VtF} that also depends on the same inputs as the Conditional Wald test. One difference is that *VtF* is nested within the *tF* approach of Lee, McCrary, Moreira, and Porter (2022), while also improving its power by allowing it to use the same information as the Anderson-Rubin approach.

It is natural to wonder whether the econometrics literature has formally determined a single best confidence interval for β. To date, it has not. Lee et al. (2023) provide numerical evidence suggesting that Conditional Wald and *VtF* may deliver narrower 95 percent confidence intervals than the Anderson-Rubin method almost all the time, if not all the time, and that they can be substantially shorter when F is relatively small. The numerical evidence is strongly suggestive, but not yet confirmed as a theorem. However, if the practical numerical implementation challenges can be overcome, there is little reason why *CW* or *VtF* would not eventually be preferred by practitioners.

In the meantime, our recommendation to practitioners is to ensure that whatever method one uses, it is a robust-to-weak-instrument method, keeping in mind that future research may well formally establish that Anderson-Rubin is dominated by an alternative procedure in terms of delivering an equally valid, but narrower confidence interval. Such a finding would not change the fact that Anderson-Rubin intervals are valid; it would just indicate that they are unnecessarily wide.

Some Do's and Don'ts

Each of the methods in the previous section will work on their own for delivering correct inference without taking a stand on the values of ρ or $E[F]$. Here, we warn of some temptations to sidestep when working with the instrumental variable model.

Do Not Form Ad-Hoc "Hybrid" Procedures

It might be tempting to combine the procedures, perhaps based on the observed value of F. For example, it may sound reasonable to say: "I'll first look at my F statistic. If it's large enough—say, close to or larger than 104.67—*tF* says I can use the usual ± 1.96 confidence interval. But if it's smaller than 104.67, I'll switch to *AR*." However, this "hybrid" procedure does not deliver correct inference. Both *tF* and *AR* are separate, prespecified plans that map what you observe in the data to conclusions about the hypothesis like "reject" or "do not reject." Combining the tests in this way inappropriately gives the researcher "two bites at the apple," so that any claims that the test is ensured to produce no more than a 5 percent rate of Type I error will be false. It may be possible to construct hybrid procedures that work, but this seemingly intuitive combination of Anderson-Rubin and *tF* does not.

Do Not Data-Mine for Large F Statistics

The following logic might be tempting: "I have a number of possible variables to use as my single instrument. I know that there are problems if the first-stage F is low, so I will set a threshold and just not consider any instrument whose first stage produces an F less than that. However, if the F statistic exceeds my threshold, I'll just go ahead and use the usual two-stage least squares t-ratio method."

The main problem is that this data-mining for an instrumental variable with a large value of F essentially *truncates* the data, which will lead to even larger distortions in inference. As noted earlier, Staiger and Stock (1997) derived the approximate, non-normal distribution of various statistics, including the t-ratio, but they were derived under the presumption that no truncation on the basis of F should occur. Discarding the cases where the F is low introduces a whole new layer of distortions that have not been extensively studied.

The Use of the "*F* Greater than Ten" Rule-of-Thumb for Testing or Confidence Intervals for the Just-Identified Model Has No Theoretical Justification

The "F at least ten" notion is something that is part of the common lore, but it is helpful to understand what are appropriate and inappropriate uses of the threshold of ten. As noted in Stock and Watson (2019), the use of the threshold of ten can be discussed meaningfully, but only in the context of the *over-identified* instrumental variable model (more than one instrument)—not for the just-identified case. In the just-identified case, when one is reporting tests or confidence intervals, the fact that a first-stage F statistic exceeds ten has no particular meaning.

The Critical Values in the Tables in Stock and Yogo (2005) Cannot Be Used to Reject Hypotheses at the 5 Percent Level or Produce 95 Percent Confidence Intervals

The critical values in Stock and Yogo (2005) are often casually cited by practitioners to indicate that their first-stage F is sufficiently strong to use the usual ± 1.96 critical values or confidence intervals. It turns out that for the just-identified case, there is only one number in the paper that is relevant for these purposes—the threshold of 16.38 (Table 5.2). Without going into the details, the proper use and interpretation of the use of this threshold is as follows. One's "confidence interval procedure" is to live with a confidence interval that is the entire real line if $F \leq 16.38$ and otherwise with the usual ± 1.96 interval. This procedure essentially has an 85 percent confidence level—not a 95 percent confidence level.[11]

If Using a Non-robust Method, Be Explicit about It

In the standard regression model, using "homoskedasticity-only" standard errors is not incorrect by itself. The qualified statement that the inferences are only valid under the assumption of homoskedasticity, an assumption that has no guarantee of being true, is itself a correct, qualified statement. What is misleading and incorrect is to use "homoskedasticity-only" standard errors while implying validity without the qualifying statement.

[11] Stock and Yogo (2005) never made the claim that using their thresholds could produce a 95 percent confidence interval. To the contrary, their analysis made clear that even ignoring the fact that "instrument strength" ($E[F]$) must be estimated, there would at least be 5 percent "distortion," meaning the ± 1.96 confidence intervals could potentially cover the true parameter as little as 90 percent. Because F is not the same thing as $E[F]$, the back-of-the-envelope coverage rate falls to 85 percent. Using (A.7) in Lee, McCrary, Moreira, and Porter (2020), a more precise coverage rate is 90.7 percent.

A similar situation applies for inference for the instrumental variable model. That is, there may be specialized situations where one is willing to assert that the degree of endogeneity ρ, for example, is small enough that the generally non-robust t-ratio procedure is valid. As pointed out in Van de Sijpe and Windmeijer (2023), Angrist and Kolesár (2024), and Lee, McCrary, Moreira, and Porter (2022), there is a direct connection between the true value of β and the true value of ρ, so being confident enough to assert limits to the magnitude of ρ is equivalent to asserting bounds on the magnitude of β. Implementing and interpreting the confidence interval that results from this non-robust inference approach requires extra care and knowledge about plausible magnitudes from the existing literature, as demonstrated by Angrist and Kolesár (2024) in three empirical examples.[12] Our main recommendation on this front is to proceed with caution, be explicit that a priori econometric assumptions are being invoked, and at any rate, it is both costless and straightforward to additionally report the results from a robust-to-weak-instrument method (like Anderson-Rubin or tF) as a comparative benchmark.

Conclusion

In the econometrics literature, it has been established since Dufour (1997) that in the just-identified instrumental variable model, the two-stage least squares t-ratio and the usual ± 1.96 confidence interval do not deliver correct inference due to the problem of weak instruments. Staiger and Stock (1997) established that the method of Anderson-Rubin does deliver robust-to-weak-instrument inference. These basic facts have yet to be fully recognized in applied research. Using a sample of 1,311 instrumental variable specifications drawn from 61 papers published in the *American Economic Review* between 2013 and 2019, Lee, McCrary, Moreira, and Porter (2022) report that about one-quarter of the specifications report nothing beyond the two-stage least squares estimate and standard error. Less than 4 percent of the specifications use the Anderson-Rubin method. Among the ten published studies that collectively contribute the 89 specifications in Figure 2, one study mentions the test of Anderson and Rubin (1949).

We suspect that at least part of what explains this gap between theory and practice is that the weak instruments literature covers many different aspects of the problem (including estimation bias, power, and optimality) and is generally quite technically dense, and therefore does not lend itself readily to accessible translation, for example, at the level of introductory econometrics texts. Just as in the case of assuming homoskedasticity when the error terms are actually clustered and

[12] For an illustration of this approach using the data from our running example of Card (1995), see www.princeton.edu/~davidlee/wp/ivcalc.html, which use equation (7) in Angrist and Kolesár (2024) and the needed inputs to the equation. These formulas are described in the Appendix to Angrist and Kolesár (2024), and in A.8.4. from the Online Appendix of Lee, McCrary, Moreira, and Porter (2022). See Section II.E. of Lee, McCrary, Moreira, and Porter (2022) for a further discussion of how this approach plays out in a sample of studies from the *American Economic Review*.

heteroskedastic, using a non-robust method requires certain assumptions about unknown parameters to be true in order for the inferences to be correct. The robust-to-weak-instrument methods—Anderson-Rubin, *tF*, Conditional Wald, and *VtF*—work for all possible values of ρ and $E[F]$.

The points raised in this article have been explained in the context of the simple (but commonly occurring) case of a single endogenous regressor and single instrument. It should be unsurprising that none of the problems are diminished when the instrumental variable model includes multiple instruments. For a deeper discussion of the issues that arise with over-identified models, we refer the interested reader to the comprehensive survey of Andrews, Stock, and Sun (2019) for a useful starting point.

■ *We are grateful to Ariel Zhao and Kevin Cao for excellent research assistance, and to Scott Ellis for reading early drafts.*

References

Anderson, T. W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46–63.

Andrews, Isaiah, and James H. Stock. 2018. "Weak Instruments and What to Do About Them." Presentation, NBER Summer Institute Methods Lectures. https://www.nber.org/sites/default/files/2020-12/NBERSI2018_Methods%20Lectures_WeakIV1-2_v4.pdf.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11: 727–53.

Angrist, Joshua, and Michal Kolesár. 2024. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." *Journal of Econometrics* 240 (2): 105398.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Cameron, A. Colin, and Pravin K. Trivedi. 2022. *Microeconometrics Using Stata*. Vol. 1–2. 2 ed. Stata Press.

Card, David. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, edited by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky, 201–21. University of Toronto Press.

Card, David. n.d. *Data for: "Using Geographic Variation in College Proximity to Estimate the Return to Schooling."* Berkeley, CA: David Card, https://davidcard.berkeley.edu/data_sets/proximity.zip (accessed June 9, 2026).

Dufour, Jean-Marie. 1997. "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models." *Econometrica* 65 (6): 1365–87.

Keane, Michael, and Timothy Neal. 2023. "Instrument Strength in IV Estimation and Inference: A Guide to Theory and Practice." *Journal of Econometrics* 235 (2): 1625–53.

Keane, Michael P., and Timothy Neal. 2024. "A Practical Guide to Weak Instruments." *Annual Review of Economics* 16: 185–212.

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter. 2020. "Valid *t*-Ratio Inference for IV." NBER Working Paper 29124.

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter. 2022. "Valid *t*-ratio Inference for IV." *American Economic Review* 112 (10): 3260–90.

Lee, David S., Justin McCrary, Marcelo J. Moreira, Jack R. Porter, and Luther Yap. 2023. "What to Do When You Can't Use '1.96' Confidence Intervals for IV." NBER Working Paper 31893.

Lee, David S., and Jack Porter. 2026. *Data and Code for: "Correct (and Incorrect) Inference with a Single Instrumental Variable: Practical Takeaways from the Weak Instruments Literature."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E249877V1.

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71 (4): 1027–48.

Moreira, Marcelo J. 2009. "Tests with Correct Size when Instruments Can Be Arbitrarily Weak." *Journal of Econometrics* 152 (2): 131–40.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–86.

Stock, James H., and Mark W. Watson. 2019. *Introduction to Econometrics*. 4th ed. Addison Wesley Longman.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, edited by Donald W. K. Andrews and James H. Stock, 80–108. Cambridge University Press.

Van de Sijpe, Nicolas, and Frank Windmeijer. 2023. "On the Power of the Conditional Likelihood Ratio and Related Tests for Weak-Instrument Robust Inference." *Journal of Econometrics* 235 (1): 82–104.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 213–240

Leniency Designs: An Operator's Manual

Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár

High-stakes decisions are often made by experts: doctors decide whether to treat patients, bail judges decide whether to release defendants before trial, patent examiners decide whether to grant patents to firms, child welfare investigators decide whether to place children in foster homes, and loan officers decide whether to approve consumer loan applications. Usually, even when groups of expert decision-makers agree in many situations, there will be close calls in which systematic disagreement emerges on the appropriate course of action: some decision-makers will tend to be systematically more lenient, in that they tend to grant some treatment more often, while others will be stricter. Furthermore, which decision-maker is assigned to a given case is often either deliberately random (to ensure fairness) or as-good-as-random, due to idiosyncrasies in the assignment process (such as rotations in shifts). A *leniency design*, also known as a judge or examiner instrument design, harnesses such exogenous variation to estimate the causal effects of these high-stakes decisions by using a measure of decision-maker leniency as an instrument for the treatment in an instrumental variables regression. Such designs have exploded in popularity in recent years. Prominent applications include Dobbie and Song (2015), Dobbie, Goldin, and Yang (2018), Dobbie, Liberman, Paravisini, and Pathania (2021), Farre-Mensa, Hegde, and Ljungqvist (2020a),

■ *Paul Goldsmith-Pinkham is Associate Professor of Finance, Yale School of Management, New Haven, Connecticut. Peter Hull is Professor of Economics, Brown University, Providence, Rhode Island. Michal Kolesár is Professor of Economics, Princeton University, Princeton, New Jersey. Their email addresses are paul.goldsmith-pinkham@yale.edu, peter_hull@brown.edu, and mkolesar@princeton.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20251480.

Autor and Houseman (2010), Doyle (2007), and Mullainathan and Obermeyer (2022) (see Table 1 of Frandsen, Lefgren, and Leslie (2023) for more examples).

Leniency designs rest on a straightforward idea: if decision-makers are randomly assigned, and if assignment only affects outcomes through the treatment decisions, dummies for assignment to different decision-makers are valid instruments. Specifically, we should be able to use the two-stage least squares estimator to estimate the treatment effect. This estimator aggregates the multiple dummy instruments into a single instrument, given by the fitted values from the first-stage regression of treatment on the assignment dummies. With no controls, the fitted values correspond simply to the sample treatment-granting propensity—or sample leniency—of each decision-maker. The treatment effect estimate is then calculated as the ratio of the sample covariance of this leniency measure with the outcome relative to its covariance with the treatment.

While intuitive, this simple idea turns out to have many practical complications. The two-stage least squares estimator works well when there are only a few decision-makers. But typically there are many, which makes the sample leniency a noisy measure of its population analog. This noise causes the two-stage least squares estimator to produce estimates that replicate some of the selection bias present in ordinary least squares estimates. What is a more robust way of aggregating the assignment dummies—a better leniency measure—that avoids such bias? Other practical questions abound: What controls should we include to ensure credible treatment effect estimates? Should we cluster standard errors, and if so, at what level? Does the fact that leniency is estimated matter for inference? When can we be confident our estimates are meaningful when treatment effects are heterogeneous?

This paper develops a step-by-step guide to leniency designs that answers these and other questions, complementing a recent review by Chyn, Frandsen, and Leslie (2025). We show how the unbiased jackknife instrumental variables estimator (UJIVE) of Kolesár (2013) is purpose-built for leniency designs, leveraging leniency measures that can ensure bias-free causal estimates even in the presence of many decision-makers or controls. We further explain how one can interpret UJIVE estimates under treatment effect heterogeneity, using the local average treatment effect framework of Imbens and Angrist (1994). UJIVE is not only useful for estimating treatment effects in this framework: building on Abadie (2002) and Kitagawa (2015), we show how it can also be used to test a key identifying assumption of average monotonicity. Finally, we show how knowledge of the decision-maker assignment process—or "design"—can guide the choice of controls and standard error calculations, and that nonclustered standard errors are often called for.

We conclude with a practical checklist for estimating treatment effects, assessing key assumptions, and probing external validity, all with a unified UJIVE approach. We illustrate this checklist with a re-analysis of Farre-Mensa, Hegde, and Ljungqvist (2020a), who use quasi-random patent examiner assignment to estimate the value of patents to startups. We confirm their overall finding that patent-granting significantly increases future patent applications, approvals, and citations on both the extensive and intensive margins, though the magnitude and precision of these estimates are sensitive to using the more robust UJIVE estimator.

Motivating Example: Estimating the Value of Patents

We start with the simplest possible version of a leniency design, with two decision-makers who are completely randomly assigned. Concretely, consider a stylized version of the setting in Farre-Mensa, Hegde, and Ljungqvist (2020a), who estimate the causal effect of patent-granting to startup firms on the later inventiveness of those firms. We observe a set of n applications, indexed by i, which are submitted by firms to the US Patent and Trademark Office. The applications are then assigned by a random coin flip to one of two examiners, s and t. The examiners decide whether to grant the patent, as indicated by the dummy variable $x_i \in \{0,1\}$. We refer to x_i as treatment, following standard causal inference lingo. An outcome y_i is then realized; for concreteness, suppose y_i measures the number of future patent applications filed by the firm over some time period (for example, within five years of submitting the application).

We want to leverage the randomness in examiner assignment to estimate the causal effect of x_i on y_i. For now, we assume this effect is the same across all firms: that is, that being granted a patent causes each firm to submit β additional patent applications, compared to being denied a patent. In practice, it is likely that the effects of patent-granting are heterogeneous (in this case, different for different firms); we allow for this possibility later. Setting aside such heterogeneity for now, we write a constant-effects outcome equation for y_i:

$$y_i = \gamma + \beta x_i + \varepsilon_i,$$

where the intercept γ represents the average number of future patent applications submitted when a firm is denied a patent. The actual number of future patent applications varies across firms due to various unobservable factors captured by the error term ε_i.

The core identification challenge here is that estimation of β by simple ordinary least squares is likely to suffer from selection bias: startups that would produce many patents in the future even if their current application were denied (because, say, they have higher research and development budgets or employ better scientists) may submit stronger applications that are more likely to be granted. This results in a positive correlation between the error term ε_i and the treatment x_i. Consequently, an ordinary least squares regression of the outcome y_i on the treatment x_i will tend to overstate the true causal effect β.

A leniency design addresses selection bias by leveraging differences in the randomly assigned examiners' tendency to grant patents, isolating variation in the treatment that is unrelated to ε_i. Suppose examiner t is "tough" (less likely to grant a patent) while examiner s is "soft" (more likely to grant a patent). In other words, the overall patent-granting rate of examiner s in the population of patent applications, which we refer to as that examiner's leniency, is larger than that of examiner t: $p_s > p_t$. Let $\hat{p}_s$ and $\hat{p}_t$ be estimates of these rates—the fraction of applications we observe each examiner granting in the data. Further, let z_i be a dummy variable for assignment to the soft examiner, so it equals 1 if application i

is handled by examiner s and 0 if it is handled by t. The estimated leniency of the examiner assigned to i is then given by $\hat{\ell}_i = \hat{p}_t + (\hat{p}_s - \hat{p}_t)z_i$. A leniency design uses estimated leniency as an instrument for patent-granting x_i. Specifically, we estimate β by an instrumental variable regression, using $\hat{\ell}_i$ to instrument for x_i in the outcome equation.

To gain some intuition as to why this instrumental variable regression can work, note that in this simplified setting—with only two randomly assigned examiners—it is equivalent to a simpler instrumental variable regression that uses the examiner assignment dummy z_i directly as an instrument. In other words, it does not matter whether we use the estimated leniency $\hat{\ell}_i$ as an instrument or the dummy z_i: the resulting estimates of β are identical. This is because the estimated leniency is just a version of z_i that is scaled by $\hat{p}_s - \hat{p}_t$ and shifted by $\hat{p}_t$, and such instrument scaling and shifting leaves estimates unchanged.

It is easy to see why the equivalent z_i-instrumented specification is valid: z_i is randomly assigned, like treatment assignments in a simple randomized controlled trial. Thus, so long as examiner assignment only affects the outcome y_i through the patent-granting treatment x_i—the usual *exclusion restriction* for instrumental variable analyses—z_i will be independent of ε_i and hence be a valid instrument for estimating β.[1] Exclusion seems reasonable to assume in this running example, unless one thought patent examiners did other things besides ruling on the patent application (such as giving advice to the firms).

Forming valid standard errors is also easy in this example. Because z_i is randomly assigned by a coin flip for each application, it is uncorrelated across applications. It follows that conventional heteroskedasticity-robust standard errors are appropriate for quantifying uncertainty in the estimate. There is no need to cluster standard errors, analogous to how clustering is not needed in simple randomized controlled trials.

Real-world leniency designs are more complicated than this stylized example, often with many decision-makers who are not assigned completely at random. Handling these features requires some modifications, as we next discuss. Still, the core equivalence between using estimated leniency as an instrument versus using examiner dummy instruments directly will be a useful guide to thinking through estimation and inference in these more complex leniency designs.

Estimation with Many Decision-Makers and Controls

The assignment of decision-makers is rarely completely random, but institutional knowledge can sometimes imply it is as-good-as-random, at least once we condition on an appropriate set of control variables. We refer to these as *necessary controls*, as they must be included in every specification leveraging exogenous

[1] Consistency of these estimates also requires a relevance condition: that z_i and x_i have nonzero correlation. Here relevance holds because we assumed that the examiners vary in their leniency, $p_s > p_t$. This means that the population first-stage coefficient from regressing x_i on z_i, given by $p_s - p_t$, is nonzero.

assignment. As we discuss more below, patent applications in the United States are first classified depending on the technology being patented. This determines which group of specialist examiners—the so-called "art unit"—will review the application. Assignment within art units is then effectively random, conditional on the set of examiners working in the art unit at the time of the assignment. Assuming this set changes slowly over time, a researcher may take art-unit-by-cohort fixed effects as the set of necessary controls.[2]

In a realistic setting with many examiners, assignment is now captured by a vector z_i of K instruments: one for each examiner, with one examiner omitted in each art unit to prevent multicollinearity. This leads to a first-stage equation, a population regression specification linking the treatment, instrument, and controls:

$$x_i = z_i'\pi + w_i'\delta + \nu_i,$$

where the vector w_i consists of art-unit-by-cohort dummies and ν_i is a regression residual. In the special case where we just have one cohort, so that w_i reduces to art unit fixed effects, the first-stage coefficients have a simple interpretation: δ_j is the overall leniency of the omitted reference examiner in art unit j, and π_k measures the difference between the leniency of examiner k and the reference examiner in k's art unit.

If we knew the first-stage coefficients π and δ, we could use the population first-stage fitted values $\ell_i = z_i'\pi + w_i'\delta$ as our leniency measure. By usual partialling-out logic (by way of the Frisch-Waugh-Lovell theorem), an instrumental variable regression using this instrument while controlling for w_i is equivalent to running an instrumental variable regression without any controls, but instrumenting with the residual from projecting the fitted values onto the covariates, $\tilde{\ell}_i = \tilde{z}_i'\pi$. Here $\tilde{z}_i$ denotes the sample residual from regressing the instrument vector z_i onto the controls w_i. This leads to the estimator:

$$\hat{\beta}^* = \frac{\frac{1}{n}\sum_i y_i\tilde{\ell}_i}{\frac{1}{n}\sum_i x_i\tilde{\ell}_i},$$

the ratio of sample covariances (because $\frac{1}{n}\sum_i \tilde{\ell}_i = 0$ by construction of the sample residuals $\tilde{\ell}_i$) between *relative leniency* $\tilde{\ell}_i$ and the outcome y_i versus the treatment x_i. We use the term relative leniency to stress that $\tilde{\ell}_i$ measures the leniency of the examiner handling application i relative to other examiners who could have handled the application. In contrast, ℓ_i is a measure of *absolute leniency*. With just one cohort, these measures take a simple form: ℓ_i is the overall patent-granting propensity

[2] One may optionally include precision controls in some specifications: pre-assignment characteristics that help explain variation in the outcome. For example, in the application below, one could include the number of independent claims in the patent application or class group fixed effects. Including these variables will soak up some variability in the error ε_i and thus help reduce the standard errors. This is analogous to including baseline characteristics as controls in regression specifications estimating effects of treatments in randomized controlled trials, while including the necessary controls is analogous to controlling for strata indicators in randomized trials where the treatment assignment probability varies across strata.

of i's examiner, while $\tilde{\ell}_i$ is the examiner's patent-granting rate minus the overall patent-granting rate of i's art unit.

The estimator $\hat{\beta}^*$ is approximately unbiased for β, because $\tilde{\ell}_i$ is a valid instrument when examiner assignment is as-good-as-random given the controls w_i.[3] In practice, however, it is infeasible to use this true relative leniency measure as an instrument, because we do not know the first-stage coefficients. A tempting alternative is to simply use least squares estimates of the first-stage coefficients, $\hat{\pi}$ and $\hat{\delta}$, in place of the unknown population values. This is exactly equivalent to using a two-stage least squares estimator that instruments with the full set of assignment dummies z_i. To see this, recall that the two-stage least squares estimator is equivalent to an instrumental variables estimator that uses a single instrument given by the sample first-stage fitted values, $z_i'\hat{\pi} + w_i'\hat{\delta}$. By the same partialling-out logic that led to the equation for $\hat{\beta}^*$, this is in turn equivalent to an instrumental variable regression without covariates that replaces the population relative leniency $\tilde{\ell}_i$ in the formula for $\hat{\beta}^*$ with the sample analog $\hat{\ell}_i = \tilde{z}_i'\hat{\pi}$ (the residual from projecting the fitted values onto the controls).

However, it turns out using the estimated relative leniency $\hat{\ell}_i$ will tend to produce biased treatment effect estimates when there are many examiners. This is the classic many-weak instrument bias problem for two-stage least squares (for example, Bekker 1994; Bound, Jaeger, and Baker 1995); it arises here because $\hat{\ell}_i$ is constructed partly from observation i's treatment status through the least squares estimate $\hat{\pi}$. With just one cohort, for instance, $\hat{\ell}_i$ is given by the fraction of patents granted by the examiner assigned to i in the sample minus the fraction of patents granted by the art unit handling the application—and both fractions are computed including the data from application i. But applicant i's treatment likely correlates with the outcome error ε_i; this is, after all, the reason to use an instrumental variable instead of an ordinary least squares regression. Thus, the estimated leniency instrument is also likely correlated with the outcome error, and this generates bias in the instrumental variable estimates in the direction of the naïve ordinary least squares estimates.

The bias from using the estimated leniency instrument can be severe, especially when examiner assignment only modestly affects the probability of treatment. A helpful rule of thumb for gauging the magnitude of the bias obtains when we assume the errors are homoskedastic; then, the two-stage least squares bias approximately equals the bias of ordinary least squares divided by $(1 - R^2) \times E[F]$, where $E[F]$ is the expectation of the first-stage F statistic for the hypothesis that the first-stage instruments are irrelevant (that is, that $\pi = 0$ in the first-stage equation), and R^2 is the population partial R-squared from adding instruments to the first-stage regression. In practice, because R^2 tends to be near zero, the magnitude of $E[F]$ tells us how much smaller two-stage least squares bias is relative to ordinary least squares (it is never larger, because $E[F]$ always exceeds one). In fact, this relationship can be used to justify the popular $F > 10$ rule of thumb proposed by Staiger and Stock (1997) for identifying weak instruments. If the expected first-stage F statistic lies below 10, then two-stage least squares bias exceeds 10 percent of least squares bias,

[3] Supplemental Appendix A provides the formal derivation of this claim, along with other results below.

so the rule of thumb can be thought of as a diagnostic for whether we are in this large bias region.[4] With many patent examiners, the first-stage F statistic can be modest even when examiner leniency explains economically meaningful variation in the treatment, because the formula for F divides by K.

Knowing that the two-stage least squares bias comes from using own treatment status to estimate the relative leniency suggests a natural bias-free alternative: instrument with a leniency estimate $\hat{\ell}_{-i}$ that leaves out observation i's own value of the treatment, thereby avoiding a mechanical correlation with ε_i. The unbiased jackknife instrumental variable estimator (UJIVE), studied in Kolesár (2013), implements this logic by setting the relative leniency estimate to $\hat{\ell}_{-i} = \tilde{z}_i'\hat{\pi}_{-i}$ where $\tilde{z}_i$ are the same instrument residuals as before and $\hat{\pi}_{-i}$ is a least-squares estimate of π that uses all observations except for i. Because this leave-one-out—or "jackknifed"—estimate is unbiased for π and, in independent and identically distributed data, independent of the data for observation i, the same arguments used to show approximate unbiasedness of the infeasible estimator $\hat{\beta}^*$ can also be used to show approximate unbiasedness of the UJIVE estimator. Moreover, UJIVE is simple to run (as shown in Supplemental Appendix A): the vector of relative leniencies can be calculated in a single step, using some matrix algebra tricks, so that we do not need to run n separate leave-out regressions to get the $\hat{\pi}_{-i}$'s.

The unbiasedness property of UJIVE hinges on the fact that it uses leave-out estimation to estimate relative leniency $\tilde{\ell}_i$ directly, accounting for controls through the residualized $\tilde{z}_i$. In contrast, leave-out estimation of absolute leniency ℓ_i tends to work poorly when the number of covariates is large. In particular, Phillips and Hale (1977) and Angrist, Imbens, and Krueger (1999) propose the jackknife instrumental variable estimator (JIVE), which constructs a leave-out estimate of the absolute leniency as an instrument in a regression with covariates; they call this estimator JIVE1; the related JIVE2 estimator has similarly poor performance with many covariates. By the same partialling-out logic as before, this is equivalent to using two-stage least squares without controls and using as an instrument the least squares residuals from a regression of the leave-out absolute leniency estimates on the controls. Even though the JIVE absolute leniency estimates are by construction uncorrelated with own treatment, the covariate adjustment reintroduces own-observation bias because the residuals depend on absolute leniency estimates for other observations—which are constructed using one's own treatment status.

To gain some intuition on this problem of classic JIVE, consider the case of just one cohort, so the covariates consist of art unit fixed effects. The JIVE absolute leniency estimate of application i is then the average leave-out patent-granting rate of the examiner assigned to i. We regress these on art unit fixed effects using ordinary least squares and take the residuals. The fitted value is the average of the leave-out patent-granting rates in the art unit assigned to i. Crucially, because

[4]Stock and Yogo (2005) construct a formal statistical test based on F for the null hypothesis that the two-stage least squares bias is large in the sense of potentially exceeding 10 percent of least squares bias, under homoskedastic errors. While the precise critical value depends on K, it is close to 10 for most values of K: see their Table 5.1.

leave-out patent-granting rates for all other cases handled by i's examiner depend on the treatment of application i, the average of the leave-out rates is the same as the overall (non-leave-out) average rate of the art unit (the average leniency of all examiners in the art unit).

Specifically, consider an examiner j working in a given art unit, who handles n_j cases $\mathcal{J}$. The sum of the leave-out rates for this examiner is $\sum_{i\in\mathcal{J}}\left(\frac{1}{n_j-1}\left(\sum_{i'\in\mathcal{J}}x_{i'}-x_i\right)\right)=\frac{1}{n_j-1}\left(n_j\sum_{i'\in\mathcal{J}}x_{i'}-\sum_{i\in\mathcal{J}}x_i\right)=\sum_{i\in\mathcal{J}}x_i$—just the overall number of cases granted by j. Consequently, summing leave-out rates for all examiners working in the art unit and dividing by the number of total cases gives the average patent-granting rate of the art unit. Thus, the residualized JIVE leniency takes the leave-out patent-granting rate of examiner assigned to i and subtracts the overall patent-granting rate of the art unit, which depends on the treatment status of application i. This reintroduces the own-observation bias of two-stage least squares that we were trying to get rid of, except that it typically runs in the opposite direction to two-stage least squares (because the own-observation component is now only in the term being subtracted off; if i is treated, this decreases JIVE's relative leniency measure). Under homoskedasticity, the JIVE bias approximately equals that of two-stage least squares multiplied by $-E[F]\times L/((E[F]-1)K-L)$, where L is the number of controls. The JIVE bias is negligible when the number of controls is much smaller than the number of instruments (in fact, UJIVE and JIVE coincide in the absence of controls and a constant). But when many controls (in our running example, art-unit-by-cohort fixed effects) are needed to ensure as-good-as-random assignment, this formula shows that the magnitude of JIVE bias can be comparable to that of two-stage least squares.

Two more comments are warranted here. First, it is common practice to report the first-stage F statistic with two-stage least squares estimates; as discussed, its expectation $E[F]$ informs two-stage least squares bias. But a researcher using UJIVE need not worry about small first-stage F statistics: the arguments for its approximate unbiasedness work even if $E[F]$ is small. In fact, UJIVE remains approximately unbiased and consistent even when the instruments are weak enough that $E[F]$ converges to one in large samples, so long as $\sqrt{K}\times\left(E[F]-1\right)$ is large (for example, Mikusheva and Sun 2022). Second, the leave-out estimation approach leveraged by UJIVE assumes independence across observations. When observations are instead clustered together—a scenario we detail when discussing the calculation of standard errors—a leave-own-cluster-out modification of UJIVE may be needed to ensure unbiasedness (Frandsen, Leslie, and McIntyre 2025; Kolesár, Min, Wang, and Zhang 2026). Intuitively, correlations between treatment x_i and errors ε_j for observations i and j in the same cluster will generally reintroduce the mechanical bias between the leave-own-observation-out UJIVE instrument and the errors, again tending to bias estimates toward ordinary least squares. Hence, for examiner designs, clustering in the data is not just a consideration for inference but also guides the choice of estimator.

Alternatives to UJIVE

While UJIVE is a natural solution to the two-stage least squares bias problem, there are other reasonable approaches. Ackerberg and Devereux (2009) propose a

clever modification of JIVE—the improved jackknife instrumental variable estimator (IJIVE)—that can greatly reduce its bias in the presence of controls by reversing the order of operations: first residualize the instruments and then compute a leniency measure based on leave-out fitted values. While reversing the order of operations does not entirely eliminate the own-observation bias (because the initial residualization is not leave-one-out), in practice IJIVE and UJIVE tend to produce similar estimates.

More recently, Chao, Swanson, and Woutersen (2023) propose an alternative to UJIVE, which they term FEJIV. We show in the Supplemental Appendix that FEJIV can be interpreted as solving for a relative leniency measure with the smallest mean squared error under homoskedasticity, subject to two constraints: that the resulting estimator is free of own-observation bias and that the leniency measure is orthogonal to the covariates. The UJIVE leniency measure, in contrast, achieves the minimal mean squared error without the latter orthogonality constraint. The orthogonality property of FEJIV is attractive in that adding a linear function of covariates to the outcome does not affect the estimate; the price for this is that FEJIV leniency is slightly noisier. The resulting estimator also tends to be computationally demanding in large datasets and imposes stronger data requirements (for example, the estimator may not exist if there are high-leverage observations).

Another approach is to bias-correct two-stage least squares, which again replaces the infeasible instrument $\tilde{\ell}_i$ in the formula for β^* with the estimated relative leniency $\hat{\ell}_i$. When two-stage least squares does this, it increases the expectation of the denominator from $\frac{1}{n}\sum_i \tilde{\ell}_i^2$ to $\frac{1}{n}\sum_i \tilde{\ell}_i^2 + K\text{var}(\nu_i)/n$ under homoskedasticity. The quantity $\sum_i \tilde{\ell}_i^2$ is the numerator of the population partial R-squared statistic; it measures the increase in the explained sum of squares from adding instruments to the first stage. Because two-stage least squares uses in-sample fitted values to estimate the predictive power of the instruments, it overstates this predictive power—exactly analogous to how the unadjusted R-squared overstates the predictive power of regression (the same issue happens in the numerator, and taking the ratio gives the rule-of-thumb bias formula for two-stage least squares). In the same way that the adjusted R-squared fixes this issue with a degrees of freedom correction, we can use degrees of freedom adjustments in both the denominator and the numerator of the two-stage least squares formula. Unfortunately, as with the adjusted R-squared formula, the resulting bias-corrected two-stage least squares estimator (due to Nagar 1959) only works under homoskedasticity.

In practice, rather than using UJIVE or its cousins that compute the leniency measures internally, it is common for researchers to first construct an "external" leniency measure and then use it as an instrument in a just-identified instrumental variable regression. This procedure is analogous to first computing first-stage fitted values and then using them as a single instrument rather than directly using the standard two-stage least squares estimator, which estimates everything in one step. Such "manual two-stage least squares" procedures are widely advised against (for example, Angrist and Pischke 2009, Chapter 4.2) and we would similarly advise

against manual leniency instrumental variable estimation.[5] Indeed, as shown above, subtle variations in how the leniency measure is exactly constructed and accounts for covariates can have potentially large consequences for the bias of the ultimate estimator. Also, just as manual two-stage least squares implementations lead to incorrect second-stage standard errors, so do manual leniency constructions. It is simpler and safer to use a one-step implementation like UJIVE, with established unbiasedness properties.[6]

Heterogeneous Treatment Effects

So far, we have assumed the treatment parameter β is constant. This assumption can of course be a strong one in practice. It means, for example, that any successful patent application raises the future innovativeness of all startups i by the same constant amount. In reality, the value of patents is likely higher for some startups than for others. Luckily, the UJIVE estimator can retain a causal interpretation even in the presence of such heterogeneous effects—that is, when patent effects β_i vary arbitrarily across firms i.[7]

First-Stage Monotonicity

The local average treatment effect theorem of Imbens and Angrist (1994), recognized by the 2021 Nobel Prize in Economic Sciences, offers a guide to the general heterogeneous-effect interpretation of leniency designs. The theorem maintains the assumption of as-good-as-random decision-maker assignment and the instrumental variable exclusion restriction, while replacing the restriction of constant treatment effects with an assumption of first-stage monotonicity. In the patent setting, monotonicity means that if we can find an application that some examiner a would grant a patent to but some other examiner b would deny, it must be that a is more lenient in *all* cases; there cannot be another application that a would deny but b would grant. In the simple example with just two examiners, this implies we can divide the universe of firms into two groups: a "complier" group with patent applications that are granted when assigned to the soft examiner s but

[5] Likewise, we would advise against interpreting the variance of a manual (or UJIVE-derived) leniency measure as a signal of the design strength as is sometimes done in practice. For one thing, noise in estimation will tend to overstate the true variability in leniency. Properly computed UJIVE standard errors, discussed below, are sufficient for gauging the design's power.

[6] One justification for such manual leniency measure constructions is that the researcher faces missing data: there is a large sample of observations for the first-stage regression, but only a smaller subset with outcomes with which to estimate effects. This scenario is easily handled with the UJIVE approach, which simply computes the leave-out estimate $\hat{\pi}_{-i}$ on the full sample. Relative to dropping the missing data, this will increase efficiency with many examiners when the true first-stage coefficients match in the samples with and without outcome data. Otherwise, dropping the missing data may be preferred.

[7] In this section only, we assume the treatment is binary. This is mostly for notational convenience, because otherwise effect heterogeneity could come both from differences across observations and from different margins of treatment response for a given observation. See Kolesár and Plagborg-Møller (2025, Appendix A.2) for an extension of the local average treatment effect theorem to the case with multi-valued or continuous treatment.

not when assigned to the tough examiner t, and a nonresponder group whose treatment status is unaffected by examiner assignment (they are always either granted or denied, regardless of which examiner handles the case). Monotonicity rules out the presence of a third group of "defier" firms that are granted their applications only if assigned to t. The local average treatment effect theorem states that, in the absence of defiers, the instrumental variable regression using an indicator variable for assignment to the soft examiner as an instrument estimates the average treatment effect for compliers. Imbens and Angrist (1994) call this a local average treatment effect, to stress that we learn nothing about treatment effects for nonresponders. In contrast, if we ran a randomized controlled trial that granted applications by a coin flip, we would learn the overall average treatment effect.

With many as-good-as-randomly assigned decision-makers, monotonicity implies there are many complier groups (one for each pair of decision-makers). An extension of the theorem shows that using the relative leniency $\tilde{\ell}_i$ as an instrument—or the unbiased leniency measure constructed by UJIVE—identifies a weighted average of complier-group specific treatment effects, weighted using the squared leniency differences of each decision-maker pair (again, Supplemental Appendix A gives the formal result). This conclusion follows because, in the population, a leniency instrumental variable is equivalent to running a series of simple instrumental variable regressions with a single binary instrument, which restricts the sample to a given decision-maker pair, then averaging these instrumental variable regressions together using squared pairwise leniency differences as weights. Because a leniency difference measures the share of compliers, leniency instrumental variables thus weights by the square of the group size. Hence, larger complier groups receive more weight. But importantly, the weighting is convex: no complier groups are given negative weights.

Weakening Monotonicity

While monotonicity is likely more palatable than assuming constant treatment effects, it is nevertheless a strong assumption. In particular, results in Vytlacil (2002) imply it is equivalent to assuming each patent examiner has the same ranking of the patent applications' merits, with examiners only differing in the cutoff they use for gauging whether an application is above the bar. The strength of this assumption for leniency designs was, in fact, first noted in the original Imbens and Angrist (1994) analysis (see their Example 2, p. 472). There are two core issues. First, in many settings like our patent example, there are multiple dimensions to decision-makers' evaluation criteria which may make it unlikely that any two decision-makers would have the same ranking. For instance, patent applications are evaluated by their perceived usefulness, novelty, and nonobviousness, and different examiners likely place different weights on these criteria. Second, even if the weights were the same, variation in skill among decision-makers can induce ranking differences due to mistakes by less-skilled decision-makers. For instance, Chan, Gentzkow, and Yu (2022) show that in the case of radiologists, variation in skill accounts for about 40 percent of variation in their leniency.

Fortunately, the usual first-stage monotonicity condition can be weakened in three ways. To see how, first consider what an instrumental variable identifies

without monotonicity. In general, an instrumental variable regression restricted to a pair of decision-makers identifies a weighted difference between complier and defier treatment effects with weights given by the fraction of compliers and defiers (because defier treatments are shifted in the opposite direction by decision-maker assignment, relative to the compliers). Because a leniency instrumental variable averages these pairwise instrumental variable regressions, it can be seen as identifying a weighted-average treatment effect for compliers and defiers, but with negative weight put on all defier groups. This is a problem for its causal interpretation, in general. For example, we risk "sign reversals": even if patent-granting, say, uniformly increases future firm productivity so that β_i is positive for all i, the leniency instrumental variable could estimate a negative effect—namely, if the treatment effect for defiers is much larger than that for compliers. But while the presence of defiers necessarily causes a negative weighting problem in the case with a single pair of examiners, there is a subtlety with multiple examiners. Because each firm appears in multiple pairwise examiner comparisons, the total weight we place on each firm corresponds to its average complier, or defier, status across all pairwise comparisons.

This logic suggests the first way the usual monotonicity assumption can be weakened: as long as no firm is a defier "on average," the weights placed on individual firms remain positive. Equivalently, one may assume that the average leniency of the examiners who would grant the firm a patent exceeds the average leniency of those who would deny it—a condition Frandsen, Lefgren, and Leslie (2023) term "average monotonicity."[8]

The second way comes from seeing that even if some weights are negative, it only presents a problem when the patent applications with negative weights systematically differ in their treatment effects. One may be willing to assume this is not the case.[9] More generally, de Chaisemartin (2017) points out that if one finds a subset of positively weighted applications whose treatment effects match those of the firms that are defiers on average, these compliers cancel out the negative treatment weights on the defiers; we can then interpret the instrumental variable as estimating a convex weighted average of treatment effects for the remaining compliers.

The third point is that even if the negatively weighted observations differ systematically, this need not bias instrumental variable estimates substantively unless there are many firms with non-negligibly negative weights and the systematic differences in treatment effects are very large. We never see two patent examiners evaluate the same case, so we cannot directly assess the common ranking assumption. But we do have some evidence in the context of judge assignment, because some court cases are assigned to a panel of judges. There Sigstad (2026) shows that while monotonicity

[8]While the statement of the condition in Frandsen, Lefgren, and Leslie (2023) is more complex, we show in the Supplemental Appendix that our formulation is equivalent. With two decision-makers, average monotonicity reduces to the usual monotonicity condition.

[9]Heckman, Urzúa, and Vytlacil (2006) call this assumption no "essential" treatment effect heterogeneity. In terms of the Roy (1951) selection model, it imposes no "selection-on-gains."

is technically often violated in judicial panels, the ranking disagreements are not severe enough to create substantial bias in leniency instrumental variable estimates.

Testing Monotonicity

Typically, leniency instrumental variable designs do not feature multiple decision-makers being assigned to the same case, precluding such direct monotonicity tests in other contexts. But we can still indirectly test the assumption. One implication of monotonicity is that the average outcomes for cases assigned to two decision-makers with the same leniency must be the same in the population, because under monotonicity, the two decision-makers' decisions must exactly match. More generally, if the leniencies are similar, the average outcomes must also be similar, because the number of observations on which the decision-makers disagree must be small. Frandsen, Lefgren, and Leslie (2023) formalize this idea in a statistical test.[10] While useful in some settings, the test has two limitations. First, for validity, Frandsen, Lefgren, and Leslie (2023) rule out a large number of decision-makers (which was the original motivation for using UJIVE), and its implementation requires bounded outcomes (ruling out, for example, looking at patent effects on future startup profits). Second, it tests the original Imbens and Angrist (1994) monotonicity condition, not the weaker average monotonicity condition that is necessary and sufficient for nonnegative weights.

To address both limitations, we propose here a test for average monotonicity—based on an earlier proposal by Kitagawa (2015)—which leverages the UJIVE estimator.[11] We develop our proposed monotonicity test by first noting, following Abadie (2002), that the UJIVE estimator can not only estimate average treatment effects for compliers; it can also characterize compliers by their baseline characteristics and potential outcomes. Formally, consider a UJIVE estimator with the same treatment, instruments, and controls as before, but with a modified outcome. Instead of y_i, we put on the left-hand side $\tilde{y}_i = v_i \times x_i$, where v_i equals some variable determined before the as-good-as-random assignment of z_i. The local average treatment effect theorem applies to this modified outcome as well, showing that under average monotonicity, the UJIVE estimator identifies a convex weighted average of treatment effects of x_i on $\tilde{y}_i$. But by construction these "effects" are just v_i, because $\tilde{y}_i$ changes by exactly this amount when x_i is counterfactually increased by one unit. Moreover, the weights that UJIVE puts on these effects are exactly the same as the ones in the original treatment effect specification. Hence, by simply replacing y_i with $\tilde{y}_i$, we can easily compute weighted averages of v_i with the UJIVE weights.

[10] Strictly speaking, both this test and the average monotonicity test described below are of the joint null hypothesis of as-good-as-random assignment, exclusion, and monotonicity, as a violation of any assumption could drive a test rejection.

[11] An alternative informal test discussed in Frandsen, Lefgren, and Leslie (2023), and used previously in the literature (for example, Dobbie and Song 2015; Dobbie, Goldsmith-Pinkham, and Yang 2017), is to check that leniency rankings of decision-makers are stable across subsamples defined by baseline characteristics. For example, subgroup regressions of treatment on relative leniency should yield a positive coefficient in each subgroup. Establishing and comparing the power of such tests with that of our proposal is an interesting area for future research.

To see why this result is useful for testing monotonicity, note that it is possible to obtain a logically invalid estimate when computing such weighted averages. If v_i is a binary variable, which can only take on values zero or one, then any convex weighted average of v_i must lie between zero and one. Such a finding would suggest something has gone wrong in the local average treatment effect theorem, namely monotonicity (assuming we are confident in as-good-as-random assignment and exclusion). More generally, we can set v_i to be an indicator that some baseline variable—or set of variables—takes on a particular set of values and check that the resulting UJIVE estimate with that $\tilde{y}_i$ as the outcome lies between zero and one. For binary outcomes, we can even use the original y_i (by setting $\tilde{y}_i = y_i \times x_i$), because then UJIVE will estimate a weighted average of treated outcomes (Abadie 2002); weighted averages of untreated outcomes can be estimated by setting $\tilde{y}_i = y_i \times (x_i - 1)$. Of course, monotonicity tests like these are straightforward to implement via UJIVE, and they inherit its approximate unbiasedness property even with many decision-makers or controls. In practice, this test will likely only detect gross violations of monotonicity: to move a given weighted average beyond the logical $[0, 1]$ bounds, we need on-average defiers to be both prevalent enough and different enough from those who are compliers on average. Relative to tests for stronger versions of monotonicity (such as in Frandsen, Lefgren, and Leslie (2023) or the Coulibaly, Hsu, Mourifié, and Wan (2025) version of Kitagawa (2015)), rejections of this test will raise more meaningful concerns for a design's validity via the potential for sign reversals.

We close this section with two additional comments. First, one might wish to use the above method to estimate average baseline characteristics of compliers even when monotonicity is not a concern. This helps evaluate another, more subtle concern with instrumental variable estimation given heterogeneous treatment effects: the external validity, or representativeness, of the estimated complier-average treatment effect. By replacing the y_i outcome with $\tilde{y}_i = v_i \times x_i$ for some baseline characteristic v_i, researchers can directly evaluate whether observable complier characteristics are representative of the broader study population.[12]

Second, a further issue can arise when the specification of controls in w_i is insufficiently flexible. To avoid omitted variable bias with instruments that are only conditionally as-good-as-randomly assigned, the covariates w_i need to enter flexibly enough so that the conditional mean of the instrument vector given w_i, $E[z_i | w_i]$, is linear in the covariates. (Kolesár (2013) shows this condition is sufficient, while Blandhol, Bonney, Mogstad, and Torgovitsky (2025) show it is necessary.) This condition holds automatically when w_i is a set of fixed effects, but if other sets of controls

[12] Another complementary way of gauging external validity is to report how complier-average treatment effects, computed by simple instrumental variable regressions restricted to a given examiner pair, vary with the pairs' leniencies. Ordering the examiners by their leniency and reporting the estimates for pairs of neighboring examiners then gives an estimate of the marginal treatment effect curve (Heckman and Vytlacil 2005). Intuitively, a flat curve suggests that effects for nonresponders are likely similar to the UJIVE estimate. However, the properties of such procedures—or their more sophisticated versions (for example, Mogstad, Santos, and Torgovitsky 2018)—have not to our knowledge been formalized in the case of many decision-makers or controls.

are needed, it may require adding interactions and higher-order terms. Further flexibility may then be needed for average monotonicity to hold. The average monotonicity condition applies to the population leniency measure based on the first-stage equation (as we show in Lemma 3 of Appendix A). If, say, there are patent examiners working in multiple art units with art-unit-specific leniencies, then a first stage that only controls for art-unit-by-cohort fixed effects additively will be misspecified. This misspecification could cause average monotonicity to fail unless we run a first stage that interacts examiner assignment indicators with art-unit-by-cohort fixed effects (like the specification considered in Angrist and Krueger 1991).

However, as discussed above, even if uninteracted control specifications could yield negative weights, they need not generate meaningful bias—and will typically yield more precise estimates. In practice, researchers can explore this potential bias-variance tradeoff by checking robustness to more flexible control specifications.

Standard Errors

Leniency designs with multiple decision-makers can also raise subtle standard error issues. To see why, we start with a benchmark: suppose we knew the relative leniency $\tilde{\ell}_i$, and hence could compute the infeasible estimator $\hat{\beta}^*$. With independent and identically distributed data, the correct standard errors are given by the square root of $\sum_i \hat{\varepsilon}_i^2 \tilde{\ell}_i^2 / \left(\sum_i \tilde{\ell}_i^2\right)^2$, where $\hat{\varepsilon}_i$ is the residual from projecting $y_i - x_i\hat{\beta}^*$ on the covariates. Because the covariances in the numerator and denominator of $\hat{\beta}^*$ are approximately normally distributed by the central limit theorem, their ratio is also approximately normal with this standard error. This follows by the "delta method," a classic result in asymptotic statistics. Standard software packages compute heteroskedasticity-robust standard errors for two-stage least squares using the feasible version of this formula, plugging in $\hat{\ell}_i$ for $\tilde{\ell}_i$ and the two-stage least squares estimate of β. These standard errors are correct when treatment effects are homogeneous and the many instrument two-stage least squares bias is negligible. Under these conditions, two-stage least squares is as efficient as the infeasible estimator $\hat{\beta}^*$.

Unfortunately, with many instruments, the same issue that causes bias in the two-stage least squares estimator also causes its standard errors to be misleadingly small. Recall from our discussion of bias-corrected two-stage least squares that when two-stage least squares plugs in $\hat{\ell}_i$ for $\tilde{\ell}_i$ in the formula for $\hat{\beta}^*$, it inflates the expectation of the denominator from $\frac{1}{n}\sum_i \tilde{\ell}_i^2$ to $\frac{1}{n}\sum_i \tilde{\ell}_i^2 + K \operatorname{var}(\nu_i)/n$ under homoskedasticity. Because the same denominator appears in the standard error formula (up to the scaling by n), overfitting shrinks standard errors too. Thus, two-stage least squares mimics not only the ordinary least squares estimate but also its precision. As Bound, Jaeger, and Baker (1995) emphasize, this approach can mask the weak instrument problem: researchers may see tight standard errors around an estimate like that of ordinary least squares and wrongly think the causal effect is precisely estimated with minimal bias.

UJIVE fixes the denominator problem by construction, but the numerator of the default heteroskedasticity-robust standard error formula still has two issues.

First, with heterogeneous treatment effects, we cannot match the infeasible estimator's precision even when instruments are strong: the correct numerator picks up a second term reflecting variation in complier treatment effects across complier groups. Conveniently, Imbens and Angrist (1994) already derived the correct numerator for the standard error in their appendix (see also Lee 2018). Second, with many examiners, a third term appears due to estimation noise in $\hat{\ell}_i$ (this is the Bekker (1994) many instrument term; we give details in Supplemental Appendix A). Luckily, if we use UJIVE along with the plug-in heterogeneity-robust formula, it turns out that this also takes care of the many-instrument term—we use this formula in our empirical illustration below.[13]

Unfortunately, there are still two potential issues with UJIVE standard errors. The first is a classic one, dating back to Anderson and Rubin (1949); it concerns the earlier-mentioned delta method argument that because the covariances in the numerator and denominator of $\hat{\beta}^*$ are each approximately normal, so is their ratio. For this argument to work well, we need the signal-to-noise ratio in the denominator (the ratio of its mean to its standard deviation) to be relatively large. To see why, suppose that there is no variation in leniency across patent examiners. Then the infeasible estimator $\hat{\beta}^*$ is a ratio of two normal distributions, both with zero mean. However, ratios of mean-zero normals follow a Cauchy distribution, which has much thicker tails than a normal distribution. Typical standard errors will not account for this. When the first stage is weak so the signal-to-noise ratio is nonzero but small, the estimator will have tails that are a bit thinner than Cauchy, but still much thicker than normal tails. For the feasible estimator, the signal-to-noise ratio scales with $\sqrt{K} \times (E[F] - 1)$; if the first stage is weak enough to make this small, we have a weak instrument problem.

A vast literature has tackled the weak instrument problem over the years, including Anderson and Rubin (1949) for the case with a fixed number of instruments and several more recent papers proposing jackknife extensions for many instruments (for example, Mikusheva and Sun 2022; Matsushita and Otsu 2024; Yap 2025). The fix turns out to be straightforward. To test that the causal effect equals a particular value β_0, we compute the residual $\hat{\varepsilon}_{i0}$ from projecting $y_i - x_i\beta_0$ onto the covariates. If the null is true, $\hat{\varepsilon}_{i0}$ should be uncorrelated with relative leniency, so we just need to check whether this correlation is statistically significant. This approach turns out to be equivalent to computing the UJIVE standard error just like before and checking whether UJIVE is significantly different from β_0, with one tweak: we use $\hat{\varepsilon}_{i0}$ instead of the UJIVE residual.[14]

[13] The plug-in formula for the standard error's numerator has an own-observation bias that makes it overshoot its target, similar to the upward bias in the denominator of two-stage least squares. By a lucky coincidence, this overshooting more than accounts for the many instrument term making the formula valid (albeit slightly conservative). Correcting the slight upward bias involves a more complicated leave-three-out procedure: see Anatolyev and Sølvsten (2023), Yap (2025), and Kolesár, Min, Wang, and Zhang (2026).

[14] This is the test proposed by Yap (2025). The test proposed by Matsushita and Otsu (2024) uses the same idea, but does not use the heterogeneity-robust formula for the standard error. As a result, it is not robust to treatment effect heterogeneity, and neither is the many-instrument robust version of the Anderson and Rubin (1949) test considered in Mikusheva and Sun (2022).

Recently, for the single-instrument case, Angrist and Kolesár (2024) give a more optimistic take on the weak instrument problem: they observe that, for the case with known leniency, the usual standard error only leads to overly optimistic inference if the correlation between $\frac{1}{n}\sum_i \varepsilon_i \tilde{\ell}_i$ (the numerator of $\hat{\beta}^* - \beta$) and $\frac{1}{n}\sum_i x_i \tilde{\ell}_i$ (the denominator) is very high; if not, the weak instruments blow up the standard error with $\hat{\varepsilon}_i$ even more than if we used Yap (2025)'s robust approach that uses $\hat{\varepsilon}_{i0}$ (which in this case coincides with the Anderson-Rubin test). This correlation parameter can be thought of as a measure of endogeneity, because it corresponds to the correlation between ε_i and x_i under homoskedasticity. But severe selection biases leading to very high endogeneity are unlikely in many applications; in such cases, the usual standard errors will not mislead. An analogous argument applies to the UJIVE estimator with many instruments and controls (as we show formally in Supplemental Appendix A): the only difference is the correlation parameter uses the UJIVE leave-out leniency measure rather than $\tilde{\ell}_i$. While this correlation parameter no longer directly maps to endogeneity due to the more complicated variance structure, it can still be bounded by placing bounds on the treatment effects. In most applications, these weak-instrument issues are moot, as the variation in leniency (as measured by $\sqrt{K} \times (E[F] - 1)$) will be substantial. But if the variation is small and a high correlation parameter cannot be ruled out, using $\hat{\varepsilon}_{i0}$ to compute standard errors with the above weak-instrument test may be prudent.

Clustering

A final complication arises when the data are not independent and identically distributed, but are clustered. This can happen for two reasons: either because the sampling is clustered (say the patent office, instead of sharing the full data, only shares application data for a randomly selected subset of filing dates, and the researcher would like to generalize to the full set of data), or because the assignment is clustered (say there is a lottery for each date, and the examiner who wins the lottery is assigned all applications filed on that date). Then, even if the relative leniency $\tilde{\ell}_i$ was known, a clustered version of the standard error formula is needed. Often, a researcher observes the full population of interest (in our application, we observe all patents in a specific time frame), so the main consideration is the assignment process.

Abadie, Athey, Imbens, and Wooldridge (2020, 2023) show that if the examiner assignment process is independent and identically distributed, one does not need to worry about the correlation patterns of the residual ε_i across i, which are what traditionally motivate the decision to cluster. What matters for the standard error is the correlation pattern of the product, $\tilde{\ell}_i \varepsilon_i$. Traditional clustering considerations treat the instruments and hence $\tilde{\ell}_i$ as fixed, or conditioned upon, in which case correlation patterns in the residual do matter for the clustering decision. Under the random instrument view, in contrast, the product will be uncorrelated whenever examiner assignment is independent across i, and we can use the same rule of thumb as in randomized experiments: cluster at the level of variation in the assignment. Furthermore, whether clustering matters for the magnitude of the standard error does not help determine whether it is needed, so clustering the standard

errors "just in case" may yield overly conservative inference. This reasoning also applies when leniency is unknown, but the number of examiners is fixed.

While these arguments have not yet been formally extended to the many examiner case, it seems natural to let the assignment process guide clustering decisions. When examiner assignment is independent across cases, robust standard errors suffice. When examiner assignment is instead clustered—such as when doctors are assigned to shifts, and a single doctor covers all patients arriving in a shift—cluster at the shift level, analogous to a clustered randomized controlled trial. Importantly, as with the simple motivating example above, this line of reasoning would never justify clustering by examiners (as is sometimes done in practice).

A Checklist for Leniency Designs

We now present a step-by-step guide to leniency designs, illustrated with the replication file from Farre-Mensa, Hegde, and Ljungqvist (2020b). Their setting examines how initial patent approval by a US Patent and Trademark Office examiner affects subsequent innovation by US startups. Our reanalysis sample contains 32,514 first-time patent applications filed after 2001, with final decisions by 2013.

Table 1 summarizes key variables. The treatment—approval of the first-time application—occurs for 65 percent of applicants. Following Farre-Mensa, Hegde, and Ljungqvist (2020a), we track several key outcomes: whether startups file additional patents (40 percent do, per the table), whether any subsequent patents were approved (30 percent were), and counts of new applications, approvals, and citations. Besides standard covariates like patent class and claim counts, we observe venture capital funding rounds prior to application. For the subset of applications also submitted abroad, we observe a proxy for application quality: European or Japanese patent office decisions. We also construct a measure of local entrepreneurship using the number of startups in each headquarters' state. As we show below, these additional observables can help gauge the plausibility of the leniency design's identifying assumptions and assess external validity. (For details of variables and calculations, see Supplemental Appendix B.) Our checklist consists of five steps; to illustrate them, we use an original R package for UJIVE, available at https://github.com/kolesarm/ManyIV.

1. Identify necessary controls for as-good-as-random assignment, and determine which estimator and standard-error procedure are appropriate given the quasi-experimental design.

A credible leniency design begins with institutional knowledge that justifies quasi-random decision-maker assignment. This justification naturally identifies any required controls. In the United States, for example, once patent applications are allocated to art units, the assignment process among individual examiners is not standardized. However, Lemley and Sampat (2012) argue, based on interviews with examiners, that the assignment process is effectively random conditional on the set of examiners working in the art unit at the patent office at the time of the assignment. Many art units use the last digit of the application serial number for

Table 1
Farre-Mensa, Hegde, and Ljungqvist (2020a) Reanalysis Sample

	Mean (1)	SD (2)	Observations (3)
Panel A. Outcomes			
Any subsequent application	0.404	0.491	32,514
Number of subsequent applications	2.339	9.594	32,514
Any subsequent approved application	0.299	0.458	32,514
Number of subsequent approved applications	1.276	6.021	32,514
Any citation to subsequent patents	0.252	0.434	32,514
Number of citations to subsequent patents	5.786	52.808	32,514
Panel B. Treatment			
Application approved	0.649	0.477	32,514
Panel C. Covariates			
Patent class group I	0.175	0.380	32,514
Patent class group II	0.301	0.459	32,514
Patent class group III	0.524	0.499	32,514
Number of independent claims in application	3.730	3.174	27,226
Number of VC rounds before application	0.124	0.577	32,514
Number of startups in HQ state	317.3	354.3	32,514
Approval by European Patent Office	0.372	0.484	3,287
Approval by Japanese Patent Office	0.400	0.490	1,206

Source: Goldsmith-Pinkham, Hull, and Kolesár (2026).
Note: This table reports means and standard deviations for the sample used in reanalyzing Farre-Mensa, Hegde, and Ljungqvist (2020a). See Supplemental Appendix B for variable definitions.

assignment. Other art units similarly use rules that imply an effectively random assignment, such as a "first-in-first-out" rule that assigns each incoming application to the first available examiner. See also the institutional discussions in Sampat and Williams (2019) and Farre-Mensa, Hegde, and Ljungqvist (2020a).

While the potential examiner set is not directly observable, we can assume the set of examiners working in a given art unit changes only slowly across time and specify the necessary controls as art-unit-by-cohort (that is, year) fixed effects. Without these necessary controls, the analysis would conflate systematic differences across art units or changes across cohorts with examiner-specific variation. From our discussion of heterogeneous effects, recall that it can be important to have sufficiently flexible controls to ensure a local average treatment effect interpretation of leniency design estimates.

Quasi-random assignment is only one piece of instrument validity in an examiner design; the second is the exclusion restriction, requiring decision-maker assignment to only affect relevant outcomes through the specified treatment. Here too, a clear argument should be made from institutional knowledge. In the patent setting example, exclusion is plausible because patent examiners make relatively narrow approval decisions and likely have no direct impact on the future innovativeness of the applying firm.

The assignment mechanism can also guide the appropriate level for clustering standard errors. Individual-level randomization in the patent setting, where patent applications are routed idiosyncratically to examiners within art unit and cohort, makes heteroskedasticity-robust standard errors appropriate. Based on our rule of thumb, clustering is not needed because each case is assigned independently. Contrast this with settings where groups of observations are assigned together: if all applications from a given month were assigned to a randomly chosen examiner, we would cluster by month. As discussed above, if clustering the standard errors is needed, the UJIVE estimator should also leave out the assigned cluster.

2. Verify balance on other observables.

Good-as-random assignment implies that the average predetermined characteristics should be equal across decision-makers, conditional on the necessary controls. We test this implication directly: for each observable characteristic, we run our UJIVE specification with that characteristic as the outcome. Significant coefficients indicate imbalance and potential threats to identification. This unified approach—using the same UJIVE specification for both balance tests and treatment effect estimation—offers important advantages over alternatives. Consider the common practice of regressing observables on leniency measures constructed from a conventional first stage. With many examiners and fixed effects, this can generate mechanical correlations that masquerade as true imbalance.[15] UJIVE avoids these finite-sample biases while maintaining the interpretability of our test: the magnitude of any imbalance directly translates to potential bias in our treatment effects. If venture capital funding had a positive coefficient in this UJIVE balance check, for example, this would imply that the randomly assigned examiners' tendency to approve patents is also positively correlated with startups' expected venture capital funding. Such a finding would raise questions about whether the examiners are truly randomly assigned, and the size of the UJIVE coefficient can help quantify sensitivity to omitted variables.

Table 2 presents balance tests for the Farre-Mensa, Hegde, and Ljungqvist (2020a) reanalysis. Each row reports a UJIVE coefficient from regressing a predetermined covariate on patent approval, instrumenting with examiner indicators and controlling for art unit-by-year fixed effects. The results support quasi-random assignment: all five coefficients are not statistically significant. Just as importantly, the magnitudes are economically negligible: coefficients are typically ten times smaller than the estimated treatment effects discussed below. Consider the venture capital variable, which might raise particular concern if certain examiners systematically reviewed applications from better-funded startups. We estimate a close-to-zero effect, suggesting no systematic differences (coefficient $= -0.024$, standard error $= 0.035$). Similar null results emerge for technical complexity (independent claims), external quality validation (European patent approval), and local entrepreneurial environment (state startup

[15] Even if one uses a leave-out leniency measure, the coefficient will still be biased due to estimation error in the leniency measure, which generates an errors-in-variables bias. Another approach to testing balance is to regress the predetermined characteristics on the full set of examiner indicators and controls, and report the joint F-test for the examiner indicators. With many examiners, however, the default F-statistic is not valid (Anatolyev and Sølvsten 2023).

Table 2
Balance Checks

	Coefficient (1)	SE (2)	Observations (3)
log(number of independent claims in application)	0.066	0.084	27,226
log(1 + number of VC rounds before application)	−0.024	0.035	32,514
log(number of startups in HQ state)	−0.273	0.143	32,514
Approval by European Patent Office	0.013	0.212	3,287
Approval by Japanese Patent Office	0.525	1.258	1,206

Source: Goldsmith-Pinkham, Hull, and Kolesár (2026).
Note: This table reports results of UJIVE regressions of each covariate in panel C of Table 1 on application approval, instrumenting with examiner indicators. All estimates control for art unit-by-year fixed effects. The reported standard errors are robust to heteroskedasticity and treatment effect heterogeneity.

density). The absence of systematic imbalance strengthens our confidence that examiner assignment generates plausibly exogenous variation in patent approval.

Balance tests can also examine post-assignment variables to detect exclusion restriction violations. These tests ask whether decision-makers affect outcomes through channels beyond the specified treatment. In the patent setting, for instance, one concern is that examiner assignment affects not only whether an application is approved but also the speed of the application review—which plausibly has independent effects on follow-up innovation outcomes by reducing applicant uncertainty. A UJIVE regression that uses a measure of review speed as the outcome (for example, months under review), again keeping the same treatment, instrument, and controls, could be used to assess the significance of such potential exclusion restriction violations: if present, the research design will not generally recover the effect of patent-granting.[16]

3. Estimate treatment effects by UJIVE and alternative estimators.

The UJIVE estimator is the preferred choice for leniency designs, as it avoids the subtle potential biases of alternative approaches when there are many decision-makers and controls. Comparing UJIVE estimates and standard errors to those from these alternative approaches—such as two-stage least squares—can illustrate these biases. As usual with instrumental variable designs, it can also be instructive to compare the UJIVE estimates to ordinary least squares estimates of the treatment effect as a way to probe the importance of selection bias.

[16] When examiner assignment affects review speed, it is still possible to recover the effect of patent-granting by using review speed as a second treatment, as done in Farre-Mensa, Hegde, and Ljungqvist (2020a) (see also Autor, Maestas, Mullen, and Strand 2015). One simply uses examiner dummies as instruments for both treatments in a UJIVE regression. This approach generally requires the effect of patent-granting to be constant and the effect of review speed to be linear and constant (or at least not be selected on, see footnote 9). A more flexible approach is to interact approval with review time, discretized to a mutually exclusive set of dummies (for example, under and over two years). This can again be estimated with a UJIVE regression. Bhuller and Sigstad (2024) give conditions under which such multiple-treatment specifications have a causal interpretation under effect heterogeneity.

Table 3 presents treatment effect estimates across four specifications. Column 1 reports our preferred UJIVE estimates using examiner indicators as instruments with art unit-by-year controls. Columns 2 and 3 implement two-stage least squares with alternative instruments: the constructed leniency measure from Farre-Mensa, Hegde, and Ljungqvist (2020a), based on examiner approval rates from earlier periods in column 2, and the full set of examiner indicators in column 3. Column 4 shows ordinary least squares results.

The UJIVE estimates reveal positive and significant effects of patent approval on subsequent innovation. Startups with approved applications are more likely to file future patents, receive future approvals, and generate citations, with significant effects on both extensive and intensive margins. The pattern of bias across estimators is instructive. Ordinary least squares produces higher coefficients for subsequent applications but lower coefficients for approvals and citations, suggesting moderate selection that operates differently across outcomes. The two-stage least squares estimates using examiner indicators (column 3) generally fall between, or close to, ordinary least squares and UJIVE, pulled toward the biased ordinary least squares results. This intermediate position confirms our theoretical prediction: with many instruments, two-stage least squares suffers from a finite-sample bias that partially reproduces the selection problem it aims to solve. The UJIVE estimates avoid this bias, providing our most credible evidence that initial patent approval causally stimulates follow-on innovation. These effects operate through multiple channels—approved firms not only attempt more patents but also succeed in having them approved and generating scientific impact through citations.[17] The two-stage least squares estimates using the Farre-Mensa, Hegde, and Ljungqvist (2020a) measure of leniency tend to be larger than the UJIVE estimates. Because their leniency construction is similar to that of JIVE, this may reflect a many-covariate bias.

Standard errors also tell an important story in Table 3. The many-instrument two-stage least squares (column 3) produces standard errors roughly three to four times smaller than UJIVE (column 1)—a difference that reflects statistical pathology rather than efficiency gains. As we have discussed, two-stage least squares standard errors and estimates are both pulled toward ordinary least squares with many examiners, creating a false sense of precision around the biased estimates. The shrinkage in standard errors occurs even when using the leave-out Farre-Mensa, Hegde, and Ljungqvist (2020a) leniency measure in column 2, because those standard errors do not reflect estimation error in the leniency measure. UJIVE avoids both failures: it delivers unbiased point estimates and standard errors that accurately capture sampling variation.

4. Test monotonicity to assess the plausibility of a LATE interpretation.

To ensure that the UJIVE estimates have a clear causal interpretation under heterogeneous effects, we also need a version of first-stage monotonicity. Because

[17] Tables B.1 and B.2 in the Supplemental Appendix show how the estimates change when we additionally control for pre-application covariates. Adding such precision controls can reduce standard errors, as footnote 2 discusses. This is not generally the case here, however, because the precision gains in the outcome equation are not sufficiently large to offset the additional first-stage noise the controls introduce.

Table 3
Treatment Effect Estimates

	UJIVE	Two-stage least squares		
	Examiner indicators	FMHL approval rate	Examiner indicators	OLS
	(1)	(2)	(3)	(4)
Any subsequent application	0.173	0.265	0.232	0.234
	(0.055)	(0.023)	(0.016)	(0.006)
log(1 + number of subsequent applications)	0.323	0.456	0.374	0.357
	(0.100)	(0.037)	(0.027)	(0.009)
Any subsequent approved application	0.259	0.250	0.240	0.223
	(0.050)	(0.020)	(0.014)	(0.005)
log(1 + number of subsequent approved applications)	0.356	0.362	0.323	0.291
	(0.081)	(0.029)	(0.021)	(0.007)
Any citation to subsequent patents	0.183	0.210	0.173	0.164
	(0.049)	(0.020)	(0.014)	(0.005)
log(1 + number of citations to subsequent patents)	0.419	0.480	0.372	0.339
	(0.125)	(0.044)	(0.033)	(0.011)

Source: Goldsmith-Pinkham, Hull, and Kolesár (2026).
Note: This table reports estimates of the effect of application approval on each outcome in panel A of Table 1. Column 1 estimates effects with UJIVE, instrumenting with examiner indicators. Columns 2 and 3 instead estimate effects with two-stage least squares, instrumenting with examiners' approval rate as computed by Farre-Mensa, Hegde, and Ljungqvist (2020a) and examiner indicators, respectively. Column 4 estimates effects by ordinary least squares. All estimates control for art-unit-by-year fixed effects. The sample size is 32,514 in all specifications. Standard errors, reported in parentheses, are robust to heteroskedasticity and treatment effect heterogeneity.

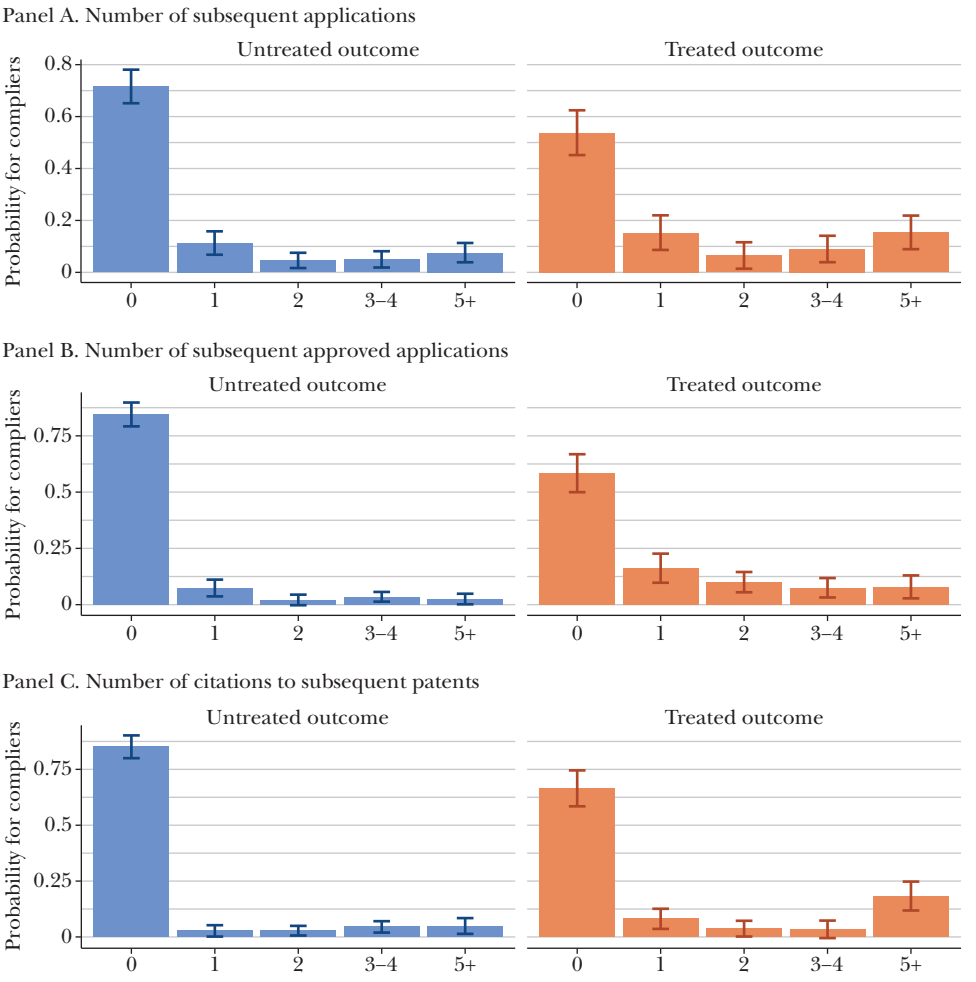
the Imbens and Angrist (1994) first-stage monotonicity condition is likely too strong in the examiner setting, we implement a UJIVE test of the weaker average monotonicity assumption: that the average leniency of the examiners who would grant the firm a patent exceeds the average leniency of those who would deny it.

Figure 1 tests average monotonicity using outcome-specific UJIVE regressions. For each plotted outcome category, we create an indicator for that value (for example, a dummy for zero subsequent applications) and interact it with treatment status as our dependent variable, maintaining examiner instruments and art-unit-by-year controls. The specification is estimated as in Tables 2 and 3. That these UJIVE estimates, and their 95 percent confidence intervals, lie within the logical bounds supports average monotonicity.

5. Characterize compliers to assess external validity.

Having found support for the design's internal validity, we next consider external validity. We estimate complier characteristics using UJIVE with modified outcomes: we regress the interaction of each covariate with the treatment indicator on the treatment itself, maintaining examiner instruments and controls. Such specifications identify weighted averages of complier characteristics for treated compliers using the same weights as our main estimates. Using one minus the treatment as the treatment variable identifies characteristics for untreated compliers; these should

Figure 1
Monotonicity Checks



Source: Goldsmith-Pinkham, Hull, and Kolesár (2026).
Note: This figure visualizes the distributions of treated and untreated outcomes for the three count outcomes in panel A of Table 1. Outcomes for treated and untreated compliers are estimated separately with UJIVE as described in the text. Vertical bars indicate 95 percent confidence intervals that are robust to heteroskedasticity and treatment effect heterogeneity.

be the same by virtue of as-good-as-random assignment. We efficiently pool these two specifications, following Angrist, Hull, and Walters (2023).[18]

Table 4 reveals that compliers resemble the full sample. Column 2 reports complier means for eight characteristics, while column 1 shows population means.

[18]For a characteristic v_i, one can show this can be done by using $\tilde{x}_i = 2x_i - 1$ as the transformed treatment variable that takes on values $\{-1, 1\}$ instead of $\{0, 1\}$. We then run a UJIVE regression of $v_i \times \tilde{x}_i$ onto $\tilde{x}_i$, maintaining the controls and examiner instruments.

Table 4
Complier Characteristics

	Sample mean (1)	Complier mean (2)	Observations (3)
Patent class group I	0.175	0.210	32,514
		(0.032)	
Patent class group II	0.301	0.307	32,514
		(0.037)	
Patent class group III	0.524	0.483	32,514
		(0.038)	
Number of independent claims in application	3.730	3.660	27,226
		(0.224)	
Number of VC rounds before application	0.124	0.158	32,514
		(0.039)	
Number of startups in HQ state	317.3	329.3	32,514
		(21.9)	
Approval by European Patent Office	0.372	0.201	3,287
		(0.097)	
Approval by Japanese Patent Office	0.400	0.430	1,206
		(0.518)	

Source: Goldsmith-Pinkham, Hull, and Kolesár (2026).
Note: This table reports means of each covariate in panel C of Table 1 for the full sample and for compliers. Complier means are estimated with UJIVE as described in the text. Standard errors, reported in parentheses, are robust to heteroskedasticity and treatment effect heterogeneity.

None of the differences are statistically significant, and all the complier estimates fall within logical bounds—again supporting average monotonicity. Compliers have similar rates of venture capital funding, comparable technical complexity in their applications, and similar representation across patent classes. This representativeness matters for interpretation: our treatment effects likely approximate average effects for the full population of first-time patent applicants, not just the subset whose outcomes depend on examiner assignment. The UJIVE estimates thus appear to broadly inform how patenting affects startup innovation.

Conclusion

Leniency designs allow researchers to estimate causal effects when expert decision-makers are as-good-as-randomly assigned and exert discretion on a high-stakes treatment. But as with many things in life and econometrics, the devil is in the details. Our review emphasizes how subtle choices in estimation and inference can have important consequences. The UJIVE estimator emerges as a natural solution to the many instrument and control problems that plague two-stage least squares estimation, while correctly estimated heteroskedasticity-robust standard errors are a natural choice when assignment operates at the individual level. Our practical checklist and reanalysis of Farre-Mensa, Hegde, and Ljungqvist (2020a) show how UJIVE can also be used for key auxiliary analyses, such as checking balance, testing average monotonicity, and characterizing compliers. We hope

this toolkit gives applied researchers more confidence when approaching leniency designs, and we encourage them to consult this manual whenever operational questions arise.

■ *We thank Joan Farre-Mensa, Brigham Frandsen, Jeffrey Kling, Emily Leslie, Alexander Ljungqvist, Timothy Taylor, and Heidi Williams for comments. Aurel Rochell provided excellent research assistance.*

References

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models." *Journal of the American Statistical Association* 97 (457): 284–92.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88 (1): 265–96.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *Quarterly Journal of Economics* 138 (1): 1–35.

Ackerberg, Daniel A., and Paul J. Devereux. 2009. "Improved JIVE Estimators for Overidentified Linear Models with and without Heteroskedasticity." *Review of Economics and Statistics* 91 (2): 351–62.

Anatolyev, Stanislav, and Mikkel Sølvsten. 2023. "Testing Many Restrictions under Heteroskedasticity." *Journal of Econometrics* 236 (1): 105473.

Anderson, Theodore W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46–63.

Angrist, Joshua, Peter Hull, and Christopher Walters. 2023. "Methods for Measuring School Effectiveness." In *Handbook of the Economics of Education*, Vol. 7, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 1–60. Elsevier.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. 1999. "Jackknife Instrumental Variables Estimation." *Journal of Applied Econometrics* 14 (1): 57–67.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.

Angrist, Joshua, and Michal Kolesár. 2024. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." *Journal of Econometrics* 240 (2): 105398.

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

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Autor, David H., and Susan N. Houseman. 2010. "Do Temporary-Help Jobs Improve Labor Market Outcomes for Low-Skilled Workers? Evidence from 'Work First.'" *American Economic Journal: Applied Economics* 2 (3): 96–128.

Autor, David H., Nicole Maestas, Kathleen J. Mullen, and Alexander Strand. 2015. "Does Delay Cause Decay? The Effect of Administrative Decision Time on the Labor Force Participation and Earnings of Disability Applicants." NBER Working Paper 20840.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62 (3): 657–81.

Bhuller, Manudeep, and Henrik Sigstad. 2024. "2SLS with Multiple Treatments." *Journal of Econometrics* 242 (1): 105785.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2025. "When Is TSLS Actually LATE?" NBER Working Paper 29709.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90 (430): 443–50.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *Quarterly Journal of Economics* 137 (2): 729–83.

Chao, John C., Norman R. Swanson, Jerry A. Hausman, Whitney K. Newey, and Tiemen Woutersen. 2012. "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments." *Econometric Theory* 28 (1): 42–86.

Chao, John C., Norman R. Swanson, and Tiemen Woutersen. 2023. "Jackknife Estimation of a Cluster-Sample IV Regression Model with Many Weak Instruments." *Journal of Econometrics* 235 (2): 1747–69.

Chyn, Eric, Brigham Frandsen, and Emily Leslie. 2025. "Examiner and Judge Designs in Economics: A Practitioner's Guide." *Journal of Economic Literature* 63 (2): 401–39.

Coulibaly, Mohamed, Yu-Chin Hsu, Ismael Mourifié, and Yuanyuan Wan. 2025. "A Sharp Test for the Judge Leniency Design." Preprint, arXiv. https://doi.org/10.48550/arXiv.2405.06156.

de Chaisemartin, Clément. 2017. "Tolerating Defiance? Local Average Treatment Effects without Monotonicity." *Quantitative Economics* 8 (2): 367–96.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review* 108 (2): 201–40.

Dobbie, Will, Paul Goldsmith-Pinkham, and Crystal S. Yang. 2017. "Consumer Bankruptcy and Financial Health." *Review of Economics and Statistics* 99 (5): 853–69.

Dobbie, Will, Andres Liberman, Daniel Paravisini, and Vikram Pathania. 2021. "Measuring Bias in Consumer Lending." *Review of Economic Studies* 88 (6): 2799–832.

Dobbie, Will, and Jae Song. 2015. "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection." *American Economic Review* 105 (3): 1272–311.

Doyle, Joseph J., Jr. 2007. "Child Protection and Child Outcomes: Measuring the Effects of Foster Care." *American Economic Review* 97 (5): 1583–610.

Evdokimov, Kirill S., and Michal Kolesár. 2018. "Inference in Instrumental Variable Regression Analysis with Heterogeneous Treatment Effects." Unpublished.

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist. 2020a. "What Is a Patent Worth? Evidence from the U.S. Patent 'Lottery.'" *Journal of Finance* 75 (2): 639–82.

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist. 2020b. *Data for: "What Is a Patent Worth? Evidence from the U.S. Patent 'Lottery.'"* Distributed by Wiley Online Library. https://doi.org/10.1111/jofi.12867.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review* 113 (1): 253–77.

Frandsen, Brigham, Emily Leslie, and Samuel McIntyre. 2025. "Cluster Jackknife Instrumental Variables Estimation." *Review of Economics and Statistics*. https://doi.org/10.1162/rest.a.263.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2026. *Data and Code for: "Leniency Designs: An Operator's Manual."* Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI. https://doi.org/10.3886/E249898V1.

Heckman, James J., Sergio Urzúa, and Edward J. Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Review of Economics and Statistics* 88 (3): 389–432.

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73 (3): 669–738.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Kitagawa, Toru. 2015. "A Test for Instrument Validity." *Econometrica* 83 (5): 2043–63.

Kolesár, Michal. 2013. "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity." Unpublished.

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens. 2015. "Identification and Inference with Many Invalid Instruments." *Journal of Business & Economic Statistics* 33 (4): 474–84.

Kolesár, Michal, Pengjin Min, Wenjie Wang, and Yichong Zhang. 2026. "Cluster-Robust Inference for Quadratic Forms." Preprint, arXiv. https://doi.org/10.48550/arXiv.2602.13537.

Kolesár, Michal, and Mikkel Plagborg-Møller. 2025. "Dynamic Causal Effects in a Nonlinear World: The Good, the Bad, and the Ugly." *Journal of Business & Economic Statistics* 43 (4): 737–54.

Lee, Seojeong. 2018. "A Consistent Variance Estimator for 2SLS When Instruments Identify Different LATEs." *Journal of Business& Economic Statistics* 36 (3): 400–410.

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

Matsushita, Yukitoshi, and Taisuke Otsu. 2024. "Jackknife Lagrange Multiplier Test with Many Weak Instruments." *Econometric Theory* 40 (2): 447–70.

Mikusheva, Anna, and Liyang Sun. 2022. "Inference with Many Weak Instruments." *Review of Economic Studies* 89 (5): 2663–86.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters." *Econometrica* 86 (5): 1589–619.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *Quarterly Journal of Economics* 137 (2): 679–727.

Nagar, Anirudh L. 1959. "The Bias and Moment Matrix of the General *k*-Class Estimators of the Parameters in Simultaneous Equations." *Econometrica* 27 (4): 575–95.

Phillips, Garry David Alan, and C. Hale. 1977. "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems." *International Economic Review* 18 (1): 219–28.

Rao, C. Radhakrishna. 1970. "Estimation of Heteroscedastic Variances in Linear Models." *Journal of the American Statistical Association* 65 (329): 161–72.

Roy, Andrew Donald. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3 (2): 135–46.

Sampat, Bhaven, and Heidi L. Williams. 2019. "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome." *American Economic Review* 109 (1): 203–36.

Sigstad, Henrik. 2026. "Monotonicity among Judges: Evidence from Judicial Panels and Consequences for Judge IV Designs." *American Economic Review* 116 (1): 189–208.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–86.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg,* edited by Donald W. K. Andrews and James H. Stock, 80–108. Cambridge University Press.

Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70 (1): 331–41.

Yap, Luther. 2025. "Inference with Many Weak Instruments and Heterogeneity." Preprint, arXiv. https://doi.org/10.48550/arXiv.2408.11193.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 241–258

Stefanie Stantcheva, 2025 Clark Medalist

James Poterba and Iván Werning

The 2025 John Bates Clark Medal of the American Economic Association was awarded to Stefanie Stantcheva, the Nathaniel Ropes Professor of Political Economy at Harvard University, for her "fundamental contributions to the field of public economics, using a wide range of tools to understand the interaction of government taxation and private behavior." Stantcheva is one of the most creative and intellectually ambitious economists of her generation. This award is a wonderful recognition of the impact of her research, which displays a distinctive ability to combine rigorous theory and empirics with a sharp sense of relevant policy questions.

She has studied some of the most heavily researched questions in public economics, such as the design of optimal income taxation, the effect of tax incentives on innovation and economic growth, and the extent to which workers migrate between jurisdictions on account of taxes, and in each case discovered new insights that advance both research and policy design. She has also pursued an exciting research program at the intersection of public economics, macroeconomics, and political economy, asking how individuals form views about public policies including, but not limited to, taxes and social insurance programs. Her most recent work investigates why some individuals develop a "zero sum" mindset, regarding public policies that benefit one group as imposing a cost on another, rather than

■ *James Poterba is the Mitsui Professor of Economics, Massachusetts Institute of Technology, and President, National Bureau of Economic Research, both in Cambridge, Massachusetts. Iván Werning is the Robert M. Solow Professor of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts. Their email addresses are poterba@mit.edu and iwerning@mit.edu.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20251483.

Stefanie Stantcheva

a view that all can benefit from some policy actions. Her work not only helps to explain why some policies are enacted in democratic nations, but also offers guidance on how to improve public understanding of them.

Stefanie Stantcheva was born in Krichim, a village outside Plovdiv, Bulgaria. Her family lived briefly in Dresden, in East Germany, around the time of German unification, while her parents were in graduate school. She moved to Paris at age six when her father was hired by a multinational French energy firm. Her early interest in economics was stimulated by watching, from a distance, the transition of the Bulgarian economy from communism to capitalism. She attended the Lycée International de Saint-Germain-en-Laye near Paris, one of the premier international schools in France. The school was founded after World War II to serve the international community associated with the then-nearby NATO headquarters.

As an undergraduate at Gonville and Caius College in Cambridge, United Kingdom, Stantcheva studied economics. She spent a semester in Cambridge, Massachusetts, as part of a Cambridge University–MIT exchange program, and enrolled in a development economics course taught by Esther Duflo. She became involved in several of Duflo's research projects and remained in contact with Duflo when she returned to the University of Cambridge for her final year of undergraduate study. She earned the top first-class degree in the Economics Tripos as well as three undergraduate awards: the Gladstone Memorial Prize for Best Undergraduate Dissertation, the Adam Smith Prize for the best undergraduate student, and the prize for the best undergraduate economics dissertation. The latter was awarded

for her study of how maternal education affects children's health, written under the supervision of Thomas Crossley and Partha Dasgupta.

Next, Stantcheva enrolled in the master's degree program at the Ecole Polytechnique (ENSAE) and the Paris School of Economics. After graduation, she continued her work with Duflo, now focusing on a project on microcredit lending in developing nations.[1] With Duflo's encouragement, she applied to the MIT Economics PhD program, where she was accepted and enrolled in 2009. Although she had planned to focus on development economics, her exposure to public finance as a first-year student convinced her that tax policy was a subject that linked together many of her interests. She also began to collaborate with Emmanuel Saez, who was visiting MIT for the 2009–2010 academic year and teaching in the graduate public economics sequence. This was the start of an important collaboration. Although we had the good fortune of serving as Stantcheva's official dissertation advisers, Saez was a key source of advice and guidance during and beyond her graduate studies.

After graduating from MIT in 2014, Stantcheva was invited to join the prestigious Harvard Society of Fellows as a junior fellow. Eight previous John Bates Clark medalists, beginning with Paul Samuelson (1947) and most recently Isaiah Andrews (2021), began their academic careers in a similar fashion.[2] Stantcheva joined the faculty of the Harvard Economics Department in 2016, received tenure in 2018, and in 2021 was appointed the Nathaniel Ropes Professor of Political Economy. In 2020, she joined a group of leading Harvard faculty members on the editorial board of the *Quarterly Journal of Economics*.

In this essay, we describe Stantcheva's research in public finance, with particular attention to four research themes that were highlighted in the AEA citation: the design of optimal income tax systems, the impact of taxation on innovation, the determinants of popular support for redistributive policies, and the role of information and other factors in contributing to public support for other policies. Table 1 provides a selected list of her publications, to which we will refer in this essay by number. We cannot do justice to all her work, and we have not tried to summarize the large related literature on the topics she has analyzed. Nevertheless, we hope that this overview will encourage readers to dive deeper and to explore Stantcheva's rich research program.[3]

Optimal Income Taxation

One important early strand of Stantcheva's research focuses on optimal income taxation, an exhaustively studied subject that highlights the efficiency-equity

[1] Chakradhar (2013) describes Stantcheva's path to graduate study at MIT.

[2] The junior fellows who went on to win the Clark Medal are: Paul Samuelson, James Tobin, Steven Levitt, Amy Finkelstein, Roland Fryer, Parag Pathak, Melissa Dell, Isaiah Andrews, and Stefanie Stantcheva.

[3] We focus exclusively on Stantcheva's research contributions. Readers interested in learning more about her approach to economic research may enjoy these interviews with Stantcheva: Cutler (2021), Henríquez (2022), Levitt (2025), and Edwards and Metcalfe (2025).

Table 1
Selected List of Stefanie Stantcheva's Papers

1. "Optimal Income Taxation with Adverse Selection in the Labor Market." 2014. *The Review of Economic Studies* 81: 1296–329.
2. "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities" (with Thomas Piketty and Emmanuel Saez). 2014. *American Economic Journal: Economic Policy* 6 (1): 230–71.
3. "How Elastic are Preferences for Redistribution: Evidence from Randomized Survey Experiments" (with llyana Kuziemko, Michael Norton, and Emmanuel Saez). 2015. *American Economic Review* 105 (4): 1478–508.
4. "Taxation and the International Mobility of Inventors" (with Ufuk Akcigit and Salome Baslandze). 2016. *American Economic Review* 106 (10): 2930–81.
5. "Generalized Social Welfare Weights for Optimal Tax Theory" (with Emmanuel Saez). 2016. *American Economic Review* 106 (1): 24–45.
6. "Optimal Taxation and Human Capital Policies over the Life Cycle." 2017. *Journal of Political Economy* 125 (6): 1931–90.
7. "Intergenerational Mobility and Support for Redistribution" (with Alberto Alesina and Edoardo Teso). 2018. *American Economic Review* 108 (2): 521–54.
8. "A Simpler Theory of Optimal Capital Taxation" (with Emmanuel Saez). 2018. *Journal of Public Economics* 162: 120–42.
9. "Disparities in Coronavirus 2019 Reported Incidence, Knowledge, and Behavior Among US Adults" (with Marcella Alsan, David Yang, and David Cutler). 2020. *JAMA Network Open.*
10. "The Polarization of Reality" (with Alberto Alesina and Armando Miano). 2020. *American Economic Review: Papers and Proceedings* 110: 324–28.
11. "Taxation and Migration: Evidence and Policy Implications" (with Henrik Kleven, Camille Landais, and Mathilde Muñoz). 2020. *Journal of Economic Perspectives* 34 (2): 119–42.
12. "Dynamic Taxation." 2020. *Annual Review of Economics* 12: 801–31.
13. "Diversity, Immigration, and Redistribution" (with Alberto Alesina). 2020. *American Economic Review: Papers and Proceedings* 110: 329–34.
14. "Understanding Tax Policy: How do People Reason?" 2021. *Quarterly Journal of Economics* 136 (4): 2309–69.
15. "Trust in Scientists in Times of Pandemic: Panel Evidence from 12 Countries" (with Yann Algan, Daniel Cohen, Eva Davoine, and Martial Foucault). 2021. *PNAS* 118 (40).
16. "Why do we not Support more Redistribution? New Explanations from Economics Research." 2021. In *Combating Inequality: Rethinking Policies to Reduce Inequality in Advanced Economies* Cambridge, MA: MIT University Press.
17. "Taxation and Innovation in the Twentieth Century" (with Ufuk Akcigit, John Grigsby, and Tom Nicholas). 2022. *Quarterly Journal of Economics* 137(1), 329–85.
18. "Taxation and Innovation: What Do We Know?" (with Ufuk Akcigit). 2022. In A. Goolsbee and B. Jones, eds, *Innovation and Public Policy,* University of Chicago Press, pp. 189–212.
19. "Eliciting People's First-Order Concerns: Text Analysis of Open-Ended Survey Questions" (with Beatrice Ferrario). 2022. *American Economic Review: Papers and Proceedings* 112: 163–69.
20. "Optimal Taxation and R&D Policies" (with Ufuk Akcigit and Douglas Hanley). 2022. *Econometrica* 90 (2), 645–84.
21. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." 2023. *Annual Review of Economics* 15 (1): 205–34.
22. "Civil Liberties in Times of Crisis" (with Marcella Alsan, Luca Braghieri, Sarah Eichmeyer, Minjeong Joyce Kim, and David Y. Yang). 2023. *American Economic Journal: Applied Economics* 15 (4): 389–421.
23. "Social Position and Fairness Views on Inequality" (with Kristoffer Balle Hvidberg and Claus Thustrup Kreiner. 2023. *The Review of Economic Studies* 90 (6): 3083–118.
24. "Immigration and Redistribution" (with Alberto Alesina and Armando Miano). 2023. *Review of Economic Studies* 90 (1): 1–39.
25. "Perceptions and Preferences for Redistribution." 2024. *Dimensions of Inequality: The IFS Deaton Review.*

(continued)

Table 1
Selected List of Stefanie Stantcheva's Papers (continued)

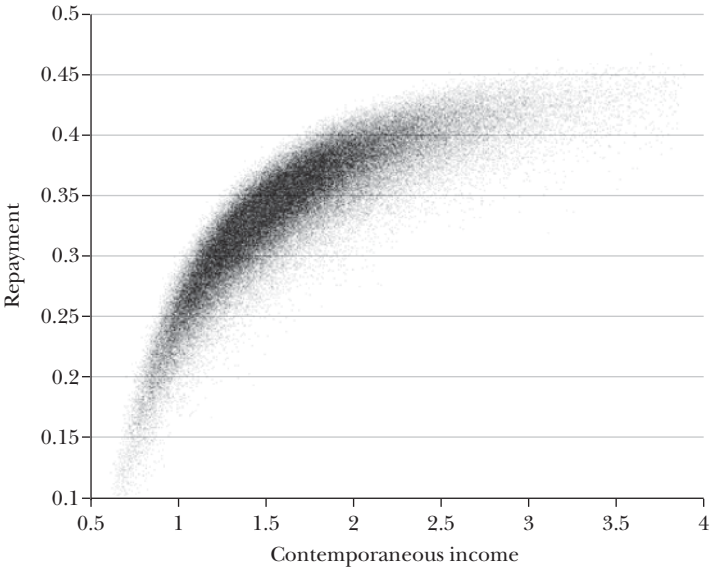
26. "Why Do We Dislike Inflation?" 2024. *Brookings Papers on Economic Activity*: 1–46.
27. "People's Understanding of Inflation" (with Alberto Binetti and Francesco Nuzzi). 2024. *Journal of Monetary Economics* 148: 103652.
28. "Fighting Climate Change: International Attitudes Toward Climate Policies" (with Antoine Dechezlepretre, , Adrien Fabre, Tobias Kruse, Bluebery Planterose, and Ana Sanchez Chico). 2025. *American Economic Review* 115 (4): 1258–300.
29. "Understanding Economic Behavior Using Open-ended Survey Data" (with Ingar Haaland, Christopher Roth, and Johannes Wohlfart). 2025. *Journal of Economic Literature* 63 (4): 1244–80.
30. "Zero-Sum Thinking and the Roots of U.S. Political Differences" (with Sahil Chinoy, Nathan Nunn, and Sandra Sequeira). 2026. *American Economic Review* 116 (3): 1052–96.

tradeoffs at the core of public finance. She has considered the tax treatment of earnings, brought new empirical evidence to bear on the debate about top marginal income tax rates, and studied optimal tax design in the presence of mobile workers.

The classic optimal income tax problem, posed by Vickrey (1945) and formalized and solved by Mirrlees (1971), is set in a static environment in which heterogeneous workers decide how many hours to work at an immutable wage. In practice, taxpayers face a more dynamic and uncertain life-cycle problem and have some control over their wage, particularly through their choice of human capital acquisition. The modern literature on optimal taxation has greatly extended the basic Mirrlees framework to incorporate many such elements, usually in isolation. In particular, Jacobs and Bovenberg (2011) allow for educational choices in deterministic two-period settings, Anderberg (2009) includes uncertainty, but abstracts from heterogeneity, and Farhi and Werning (2013) analyze a general life-cycle model with heterogeneity and uncertainty, but abstract from human capital investments.

With her novel contributions, Stantcheva has brought new richness to considerations of the broader economic and policy environment in which governments try to raise revenue. For example, many governments simultaneously encourage educational investment by subsidizing student loans, while also levying progressive taxes on labor income that discourage individuals from foregoing earnings while their wages are low in order to acquire skills that will raise their wage. Stantcheva [6] tackles the problem of jointly designing income taxation, educational subsidies, and income-contingent student loans, applying mechanism design tools in a rich framework that combines heterogeneous ability, stochastic returns, and a full life-cycle structure. She constructs an elegant dynamic model of life-cycle behavior, with much greater richness than previous studies, and characterizes both the optimal lifetime income tax schedule and the optimal set of age-dependent government subsidies for human capital investment. She incorporates uncertainty in the rate of return to education in the form of shocks to these returns over many periods, while also allowing the government to deploy a rich set of dynamic tax instruments that can condition taxes each period on observable information in the current as well as past periods.

Figure 1
Relationship Between Student Loan Repayment and Current Income for Simulated Earnings Trajectories Under Optimal Repayment Function



Source: Figure 7A in Stantcheva, Stefanie. 2017. "Optimal Taxation and Human Capital Policies over the Life Cycle." *Journal of Political Economy* 125 (6): 1931–90.

Stantcheva proves that unless the income tax schedule can condition a taxpayer's liability on educational investment, the voluntary nature of this investment reduces the optimal degree of tax progressivity. She also finds that from an efficiency perspective, distinct tax schedules and student loan programs are dominated by income-contingent loans, which can implement the optimum. In calibrated simulations of her model, she finds that the optimal loan repayment rate rises rapidly with the borrower's income, implying an important element of insurance. This is shown in Figure 1, reproduced from Figure 7A in [6], which plots repayment (vertical) against borrower income (horizontal) for simulated income outcomes under an optimal repayment schedule. The vertical dispersion of the repayment rates for each income value illustrates the historically contingent nature of the repayments: different histories of earnings shocks map into different repayments. The graph also shows, however, the upward slope of the repayment in the level of the borrower's income.

The simulations also show that much of the welfare gain of the optimal policy can be captured with a much simpler policy that involves linear income taxes, linear human capital subsidies, and linear taxes on capital income, all of which are age-specific. Allowing loan repayment rates to vary by age, for example, captures much of the welfare gain of income-related repayments. Overall, this study is both a major

technical achievement in applying mechanism design tools to a rich optimal income tax problem, as well as a foundational contribution to the design of public policies toward human capital acquisition.

Another potential complication for tax policy is that a social planner tasked with designing the optimal income tax must consider not only how workers will adjust their behavior in response to the after-tax returns to working, but also how firms will adjust the structure of their labor contracts in response to changes in the tax code. As a result, the optimal income tax is affected by asymmetric information within firms that prevents them from directly observing the productivity of their workers. In [1], Stantcheva studies the interesting and complex problem that arises when there are two layers of optimal contracting: firms must design and implement a pay structure, and the government must collect revenue by taxing the resulting pay of workers. When firms have imperfect information, their optimal compensation structure may exhibit differences between a worker's wage and marginal revenue product. Firms typically design contracts that induce workers to supply more labor than they might choose to supply in the absence of asymmetric information. The compensation structures for many high-skill employees such as senior executives and professionals who engage in team production, or lawyers and researchers, may fit this framework. In this setting, a tax increase can lead firms to modify their compensation structure to preserve these incentives for high-skill workers, complicating both the analysis of the efficiency and incidence of income taxation. This study links the theory of compensation design within organizations and optimal income tax theory.

One surprising result from this study is that introducing adverse selection and separating labor contracts into an economy without them improves welfare. While this finding may seem counterintuitive, the logic is that if the planner correctly internalizes firm contracts, these contracts enhance the available set of tools relative to the standard Mirrleesian optimal tax problem. Firms effectively become the government's partners in helping to screen workers more effectively. This research also finds, however, that if the government undertakes a standard analysis based on the Mirrlees model and fails to internalize the presence of labor contracts, then the usual optimal income tax based on sufficient statistics may lead to a suboptimal outcome.

In a multi-jurisdiction setting, another potential consequence of tax policy is distortion of the location decisions of mobile workers. Henrik Kleven, Camille Landais, Mathilde Munoz, and Stantcheva [11] point out that the elasticity of worker migration with respect to net-of-tax wages in different jurisdictions is likely to vary across worker type, and to be highest in occupations with relatively little location-specific human capital. They explore the welfare-maximizing tax rate on mobile workers when jurisdictions choose their tax rates independently, and when they are able to coordinate as part of a federation. Some countries and smaller jurisdictions have tried to capitalize on worker mobility by offering lower tax schedules to immigrants. Stantcheva and her coauthors in [11] consider the revenue and welfare effects of preferential tax regimes for foreign workers and find that

estimated elasticities of immigration can be large enough—for example in the case of Denmark—to justify preferential regimes as revenue-maximizing policies. However, they also observe that the cross-border mobility elasticity may depend on many factors, including national amenities, the degree of tax coordination across jurisdictions, and other government policies that affect cross-border population movements, and that generalization from one country to another may not be a reasonable guide for policy design.

Income tax rates can also affect the level of effort that employees devote to bargaining, and therefore their pre-tax compensation structure. This consideration becomes especially salient if the incomes of the highest earners reflect in part rent-sharing between firms and employees. Stantcheva's widely-cited study with Piketty and Saez [2] explores how corporate rents affect optimal progressivity. When top tax rates are very high, the incentive to push for an outsized executive compensation package is modest, but when top tax rates are low, actions that may result in higher pre-tax pay are more attractive. This study investigates whether the decline in top US marginal tax rates associated with the Tax Reform Act of 1986 induced those who have a role in negotiating their pay—sports stars, chief executive officers, and some others—to devote more effort to bargaining. This rent-shifting narrative is an alternative to the suggestion that the rise of earnings at the top of the distribution is the result of greater productivity for these workers in an increasingly global economy, and this paper's analysis is often cited in policy discussions of income tax progressivity.

A related study with Saez [5] examines the philosophical underpinnings of social preferences that provide the foundation for the modern theory of optimal taxation. It explores how generalized social welfare weights, derived from social justice principles, can be used to aggregate the gains and losses that different individuals will face as a result of a tax reform. It draws attention to insights from social choice theory that have not been widely discussed in public finance. In addition, it provides a concise framework in which to nest utilitarianism and its alternatives as starting points for tax analysis.

In another study with Saez [8], Stantcheva uses a wealth-in-the-utility-function model to derive optimal capital income tax rates and sufficient statistics for their implementation. The novelty of this study is its economic framework: most previous research on optimal capital taxation has used either an overlapping-generations or infinite horizon model in which wealth is valuable only for the stream of future consumption it provides. In [12], Stantcheva reviews three frameworks for the analysis of capital taxation: the dynamic Mirrlees setting, in which capital income is only taxed when doing so improves incentives to work; the dynamic Ramsey setting, in which capital taxes typically emerge as part of an optimal tax program when age-dependent labor income taxes are not possible; and the "sufficient statistics" setting in which equity-efficiency tradeoffs are analyzed in a wealth-in-the-utility model along the lines of [8]. This synthesis makes clear that the question "how should capital income be taxed?" remains unresolved, as the three frameworks give different answers and derive the optimal capital tax rate based on different economic considerations.

Taxation, Innovators, and Innovation

The effect of taxation on economic growth is one of the central issues in public economics. Both corporate and personal income taxes affect incentives for firms to invest in research and development and for individuals to pursue careers linked to innovation. Stantcheva has contributed to this important topic by bringing new identification strategies and long spans of historical data to the task of measuring the impact of taxation on innovation.

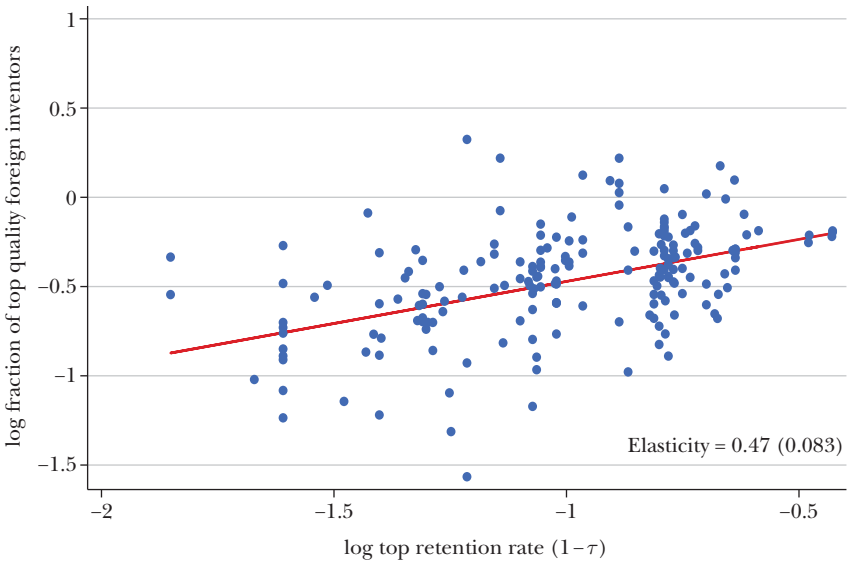
Encouraging the immigration of global research stars—a class of mobile worker whose location may be affected by taxation—is one way that a nation's tax code can promote cutting-edge innovation. In joint work with Ufuk Akcigit and Salome Baslandze [4], Stantcheva studies how individual income tax rates affect the location choices of innovators. This project draws on international patent registries and creates "location trajectories" for research stars.[4] If inventors learn about their productivity over time, those who are more successful and expect to earn more income in the future will have greater incentives to move to low-tax jurisdictions. Indeed, these incentives matter: successful inventors are more likely to migrate to countries with low top marginal tax rates than their less successful counterparts. Figure 2 (reproduced from Figure 4C in [4]) shows this relationship by plotting the share of the most-productive inventors in a country who are foreign-born against the "keep ratio," one minus the country's top marginal tax rate. The cross-country data suggest an elasticity of about 0.47 as talented investors gravitate toward low-tax jurisdictions.

Other examples of country-specific tax changes suggest even greater behavioral responses. Figure 3 (Figure 8A in [4], as reproduced in [11]), shows the evolution of the share of foreign-born inventors among the most productive inventors in the United States before and after the Tax Reform Act of 1986, which lowered top marginal tax rates from 50 to 28 percent. After assessing multiple sources of evidence, this study concludes that a 1 percent increase in the keep ratio can generate as much as a 1 percent increase in the number of top-tier foreign inventors in the country. By comparison, the sensitivity of domestic inventors is much smaller—only 0.03 percent. The findings indicate that highly progressive taxes may induce inventor emigration, a particular concern if these inventors generate positive externalities for other researchers in their country.

A range of other tax instruments might also affect the incentives of research and development (R&D) investors, along with the innovators who conduct research. With Akcigit and Douglas Hanley [20], Stantcheva shows that when firms exhibit heterogeneous R&D productivity, one-size-fits-all policies that incentivize research spending—such as an R&D tax credit at a constant rate across firms—are not optimal. For some parameterizations of the complementarity between unobserved productivity, unobserved R&D inputs, and observed R&D inputs, optimal policies can yield substantially greater research output at lower cost than simpler

[4]Kleven, Landais, and Saez (2013) use a similar methodology to track the cross-border migration of European football stars, another group that is potentially sensitive to progressive income tax schedules.

Figure 2
Foreign Fraction of High-Productivity Inventors in a Country as a Function of the Country's Top Marginal Income Tax Rate

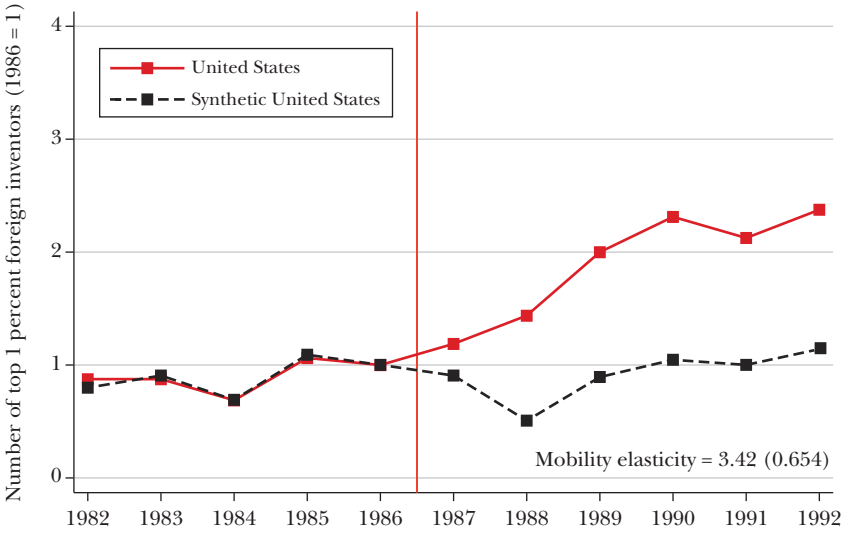


Source: Figure 4C in Akcigit, Ufuk, Salomé Baslandze, and Stefanie Stantcheva. 2016. "Taxation and the International Mobility of Inventors." *American Economic Review* 106 (10): 2930–81.

policies with constant incentives. The paper uses COMPUSTAT data on US firms, merged to patent grants, to estimate the degree of cross-firm heterogeneity. When this estimate is filtered through the optimal subsidy framework, there appear to be substantial gains from pursuing the optimal policy. While optimal policies are complex, most of the welfare gains can often be obtained with relatively simple policies such as a linear corporate income tax with a rate that declines in firm profits and a nonlinear R&D subsidy that provides a smaller subsidy for firms with the largest R&D outlays. The firms that are most productive at conducting research generate high profits and are therefore incentivized to conduct research by their low marginal tax rate, while the lower R&D subsidy at high R&D spending levels phases down the subsidy for firms spending a lot on R&D, which may be undertaking less productive marginal projects than firms that are able to achieve high profits with a smaller R&D outlay. This paper makes two significant contributions: bringing the tools of mechanism design to bear on the question of optimal R&D subsidy design and producing empirical evidence on the gains from implementing optimal subsidies that may spur greater policy attention to subsidy design.

The full range of tax policies that may affect innovation includes corporate tax rates, individual income tax rates, and research and development subsidies. Stantcheva's study with Akcigit, John Grigsby, and Tom Nicholas [17] is an empirical tour de force that analyzes data on patenting activity since 1920 and considers how

Figure 3
Share of Most Productive US Inventors Who Are Foreign-Born



Source: Figure 8A in Akcigit, Ufuk, Salomé Baslandze, and Stefanie Stantcheva. 2016. "Taxation and the International Mobility of Inventors." *American Economic Review* 106 (10): 2930–81, as reproduced in Kleven, Henrik, Camille Landais, Mathilde Muñoz, and Stefanie Stantcheva. 2020. "Taxation and Migration: Evidence and Policy Implications." *Journal of Economic Perspectives* 34 (2): 119–42.

federal- as well as state-level corporate and personal income taxes affect it. This study provides the most compelling evidence to date of a link between tax rates and innovation. State-level corporate and individual income taxes are both negatively associated with innovation within the state, both in terms of the number of inventors and new patents. A 1 percent decline in the net-of-personal income tax take-home payment for an inventor reduces the annual number of patents by about 0.8, while a 1 percent drop in the net-of-corporate income tax of the firm reduces patent activity by about 0.6 percent. There is some evidence that tax effects are attenuated in "innovation clusters" such as Silicon Valley.

There are many margins of behavioral response to taxation, and many channels through which tax changes affect research and development and innovation. Stantcheva and Akcigit [18] review the mechanisms by which general tax policies, such as individual and corporate income tax rates, and R&D-specific tax policies, such as R&D tax credits or reduced rates of tax on income from patents or other innovation-related income streams, affect the level and location of innovative activity. They argue that tax rate changes are an under-exploited source of variation in the effective price of conducting research projects. One could teach an entire module of a graduate public economics course on "public policies and R&D" using the framework developed in this study, alongside the empirical analysis in Stantcheva's other papers.

Political Economy of Redistribution

Stantcheva's innovative research in political economy uses survey-based methods to tackle two key questions: What factors underpin the public's policy preferences? How do members of the public frame and analyze questions of policy design? Of course, surveys on beliefs and attitudes towards public policies are widespread, not just in economics but also in adjacent disciplines such as political science and sociology. However, the novelty of her approach lies in constructing surveys that connect closely to the economic theories and data that bear on policy questions. She asks not only what people believe, but also what information they draw on in forming their views and what mental framework they use to evaluate policy alternatives. By integrating randomized controlled trials with traditional survey methods, Stantcheva studies how policy preferences are affected by exposure to new information or new conceptual insights, as well as how these responses align with economic modeling.

This research program began as an attempt to understand why so many individuals who would seem to be better off with a higher level of redistribution were choosing not to vote for higher and more progressive taxes coupled with more generous transfer programs. Does lack of support for redistribution come from noneconomic considerations, or from lack of information and understanding about the economic benefits that redistributive policies might generate? Stantcheva's study with Ilyana Kuziemko, Michael Norton, and Saez [3] uses online surveys to investigate how quantitative views about the extent of income and wealth inequality affect policy preferences. When survey participants learned where their household income fell in the current US income distribution, and where it would have fallen several decades earlier when inequality was lower, many expressed greater concern about economic inequality as a social problem. However, very few respondents then changed their views about the optimal top income tax rate and other redistributive policies. The only policy for which the informational intervention changed reported preferences was the estate tax. Many respondents were unwilling to support more progressive income tax rates or higher minimum wages because they were unconvinced that new policies would succeed in reducing the degree of inequality. Showing respondents data on inequality both strengthens interest in reforms that could reduce inequality, but also raises questions along the lines of "how could the government have allowed this to happen?" and reduces confidence in the effect of progressive redistribution.

In [14], Stantcheva explores what respondents understand about the economic effects of taxation. She asks about preferences regarding both income and estate taxes, then exposes participants to one of three different educational treatments. The first, "Redistribution," focused on the distributional effects of potential tax reforms on high- and low-income households. The second, "Efficiency," emphasizes the distortions associated with higher marginal tax rates. The third, labeled "Economist," blends "Redistribution" and "Efficiency" to give a sense of the tradeoffs associated with various tax reforms. The "Redistribution" and "Economist"

treatments raised support for more progressive taxes, while the "Efficiency" treatment had no statistically significant impact. There are also wide disparities in beliefs between respondents who self-identify as belonging to different political parties.[5] Democrats on average think taxes are lower than they actually are and than Republicans think they are, and they believe that the distortionary effects of higher taxes are small.

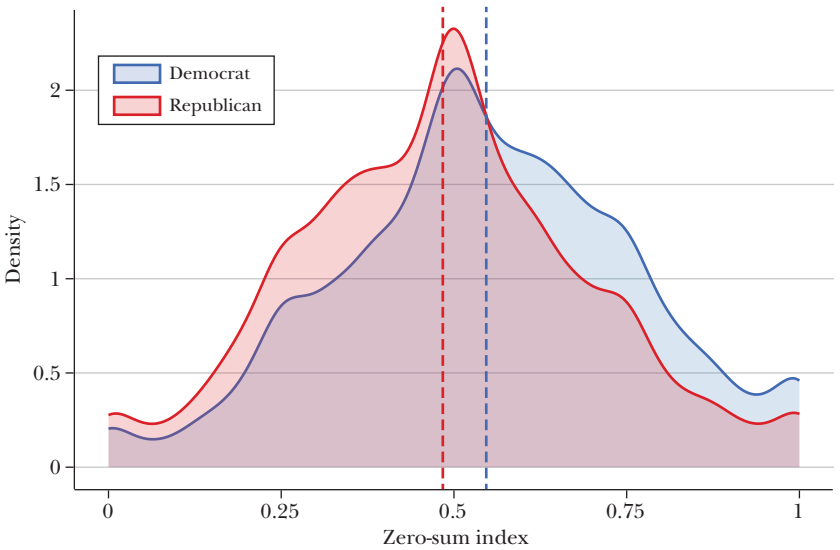
One possible explanation for limited public support for redistribution, analyzed by Luttmer (2001) and many others, is that individuals believe that the recipients of transfers will be from ethnic groups other than their own or will be immigrants rather than natives. In a joint paper with Alesina and Armando Miano [24], Stantcheva presents new evidence on how natives in six OECD countries perceive immigrants, and on the way anti-immigrant sentiment translates to anti-redistribution sentiment. Most respondents significantly overestimate the number of immigrants in their country and underestimate the degree of similarity between the immigrant and native population. When respondents are prompted to think about immigrants before they respond to questions about redistribution, they are less supportive of transfer programs. Educational interventions designed to address misperceptions of the size and composition of the immigrant population had very little effect on support for redistribution, suggesting that the narratives underlying respondent beliefs are deeply embedded. Building on these empirical findings, Alesina and Stantcheva [13] develop a conceptual framework to explain the relationship between beliefs about immigrants or minorities and support for redistribution.

In addition to documenting differences in survey respondents' perceptions of basic facts about the economy, Stantcheva also explores differences in the frameworks they bring to analyzing policy questions. Sahil Chinoy, Nathan Nunn, Sandra Sequeira, and Stantcheva [30] ask more than 20,000 US residents whether they believe that policy-induced gains for one population subgroup, from a range of policies, typically come at the expense of another group—that is, a fixed-size-pie assumption—or whether gains for a subgroup are typically associated with overall expansion of economic rewards. They find that several individual attributes are correlated with a zero-sum approach to economic policy. Immigrants and the children of immigrants, for example, are less likely to embrace zero-sum thinking. Individuals from families that have experienced relatively little upward economic mobility, and descendants of enslaved persons, are more likely to have a zero-sum view.[6] Respondents who identify their party affiliation as strongly or moderately Republican are

[5]Alesina, Miano, and Stantcheva [10] provide further evidence of partisan divide. The average survey respondent who self-identifies as conservative believes that the probability a child born into a family in the bottom quintile of the income distribution will remain there is 30 percent, compared with 37 percent for liberals. The actual probability, based on analysis of tax return data, is 33 percent. There are also disparities in the share of income that respondents believe accrues to the top 1 percent of earners.

[6]In related research that takes a cross-national perspective and explores issues raised by Alesina and Angeletos (2005) and others, Alesina, Stantcheva, and Edoardo Teso [7] find that on average, individuals in the United States are more optimistic about the prospects for intergenerational mobility and less

Figure 4
Comparison of Zero-Sum Beliefs Between Democratic and Republican Respondents



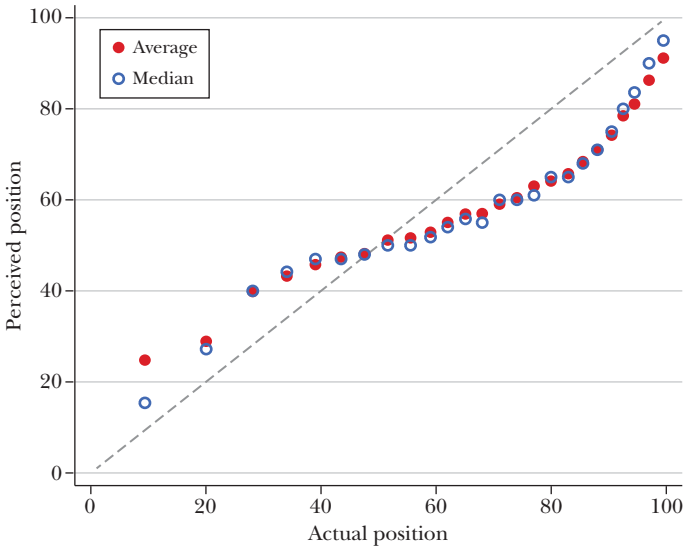
Source: Figure 3 in Chinoy, Sahil, Nathan Nunn, Sandra Sequeira, and Stefanie Stantcheva. 2026. "Zero-Sum Thinking and the Roots of US Political Differences." *American Economic Review* 116 (3): 1052–96.

less likely to subscribe to zero-sum thinking than those who identify as strongly or moderately Democratic, as Figure 4 (Figure 3 in [30]) shows. The research also provides a range of other insights into the correlation between respondent policy views and the zero-sum view. Those who are higher on the zero-sum index are more likely to support government redistribution of economic resources, affirmative action, and restrictive immigration policies.

The foregoing studies demonstrate the remarkable capacity of cleverly designed surveys to elicit information about the determinants of policy preferences, but survey evidence still has limitations. When measuring self-reported respondent attributes, for example, it is not as reliable as administrative data. For this methodology, the holy grail is to link administrative and survey data. In [23], Kristoffer Hvidberg, Claus Kreiner, and Stantcheva draw on Danish administrative data on respondents' income and other attributes to study how views about the income distribution and preferences for reducing inequality depend on an individual's taxable earnings. Figure 5, a reproduction of Figure 3 in [23], shows that Danes with low actual incomes tend to overestimate their position in the income distribution—on average those near the 10th percentile of the actual distribution report

supportive of redistributive policies than individuals in four other nations (France, Italy, Sweden, and the United Kingdom).

Figure 5
Perceived (Vertical) Versus Actual (Horizontal) Income Rank, Danish Households



Source: Figure 3 in Hvidberg, Kristoffer B., Claus T. Kreiner, and Stefanie Stantcheva. 2023. "Social Positions and Fairness Views on Inequality." *Review of Economic Studies* 90 (6): 3083–118.

themselves at the 30th, although the median response is closer to the 15th—while those near the 80th percentile of actual incomes think they are closer to the 70th. Views about the extent to which income differences are unfair appear to depend on the setting for these disparities. When differences are within well-defined groups, such as firms or groups defined by similar educational attainment, they are viewed as less fair than when they are in the broader economy.

In [16] and [25], Stantcheva draws together her findings on political support for redistribution. One lesson is that providing information—for example, data on the shape of the income distribution or the degree of intergenerational economic mobility—does not substantially alter policy preferences. Moreover, there is limited understanding of the mechanisms by which public policies might affect the distribution of resources. She concludes that there is a role for economic education in helping the public to connect the dots from policy actions to economic outcomes.

Public Understanding of Other Issues: Inflation, Pandemic Response, and Climate Change

Stantcheva has applied randomized trials within surveys to study a number of issues besides redistributive policy, including inflation, the impact of the COVID-19 pandemic on policy support, and public perceptions of climate change

and climate policy. In 2024, she fielded a survey on views about inflation and the economic policies that affect it. In [26], she reports that on average, Republicans, women, and Black people reported higher perceived rates of inflation than others. Self-identified Republicans were twice as likely (24 versus 12 percent) as Democrats to blame President Biden for the inflation increase, while the pattern was flipped (8 versus 13 percent) for greed as an explanation. Alberto Binetti, Francesco Nuzzi, and Stantcheva [27] collect additional information about the way households react to inflation. Thirty-five percent of respondents point to the increased complexity of daily decisions that results from higher inflation as an important cost, and overall inflation was viewed as a greater economic problem than high unemployment.

In April 2020, Marcella Alsan, Stantcheva, David Yang, and David Cutler [9] surveyed nearly 5,200 US adults and found significant differences across age and racial groups in knowledge and beliefs about COVID-19, with individuals over the age of 55 having the most knowledge. In a related study of COVID avoidance behavior, Yann Algan, Daniel Cohen, Eva Davoine, Martial Foucault, and Stantcheva [15] find that the public's trust in scientists is the most important determinant of compliance with non-pharmaceutical interventions. Alsan, Luca Braghieri, Sarah Eichmeyer, Minjeong Kim, Stantcheva, and Yang [22] deploy surveys in 15 countries and find that in most countries, the public, and in particular those in high-income strata, are prepared to relinquish civil rights to improve health security.

In a study of how information affects support for climate-related policies, Antoine Dechezlepretre, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stantcheva [28] find that informational interventions that provide data on the potential effects of climate change have relatively little impact, but explaining how policies affect behavior—for example by outlining how a tax on fossil fuels reduces gasoline demand from households and firms—raises support for such policies.

Stantcheva's survey-based research has attracted attention from other researchers, as well as policy analysts and the broader public, because it creatively frames important questions and develops novel ways of understanding how individuals form their views on policy. The Social Economics Lab at Harvard, which she founded, is at the forefront of this research. Stantcheva [21] reviews a number of lessons learned with regard to survey design. The frontier in this research area is constantly changing. For example, recent advances in large language models make it possible to extract more information from unstructured survey answers than was previously possible, as Ingar Haaland, Christopher Roth, Stantcheva, and Johannes Wohlfart [29] and Beatrice Ferrario and Stantcheva [19] explain.

Professional Accomplishments

The AEA Honors and Awards Committee was not the first body to recognize and honor Stantcheva's research contributions. In the year when she completed her PhD, she participated in the *Review of Economic Studies* tour. She has received a Sloan

Research Fellowship, a Guggenheim Fellowship, and an Andrew Carnegie Fellowship, a trifecta of awards from three of the leading private foundations that support economists and their research. She was awarded an NSF CAREER Award. In 2019, she was named the outstanding young French economist by *Le Monde*.

In 2020, she received the Elaine Bennett Research Prize from the AEA Committee on the Status of Women in the Economics Profession. She is an elected Fellow of the Econometric Society and the American Academy of Arts and Sciences, and a Research Associate at the National Bureau of Economic Research, where she is affiliated with the Economic Fluctuations and Growth, Political Economy, and Public Economics programs. Between 2018 and 2024, she served as a member of France's Conseil d'Analyse Économique; she is currently a member of the Cercle des Économistes.

Conclusion

Stantcheva's research sheds new light on some of the longest-standing issues in public economics and also pushes the boundaries of the field. She has framed new questions at the intersection between political economy and public economics and developed innovative research strategies for studying them. Her work has not only affected the direction of research within public economics and the economics profession more broadly, but has also attracted attention from policymakers and others who seek to understand how the public perceives and analyzes government policies.

Beyond her research contributions, Stantcheva is an active contributor to both the academic economics community and to economic policy debate. She is a standout teacher of both graduate and undergraduate students, a sought-after dissertation advisor, a mentor for the many students who have been affiliated with the Social Economics Lab, and a painstaking editor of the *Quarterly Journal of Economics*. Somehow, she also finds time to participate in policy discussions and to translate her research findings for politicians and public policy experts. Her professional trajectory provides a vivid demonstration of the many ways in which an economist can inform and improve public policy.

■ *The authors thank Jeffrey Kling, Jonathan Parker, Stefanie Stantcheva, Timothy Taylor, and Heidi Williams for comments on an earlier draft.*

References

Alesina, Alberto, and George-Marios Angeletos. 2005. "Fairness and Redistribution." *American Economic Review* 95 (4): 960–80.

Anderberg, Dan. 2009."Optimal Policy and the Risk Properties of Human Capital Reconsidered." *Journal of Public Economics* 93 (9–10): 1017–26.

Bovenberg, A. Lans, and Bas Jacobs. 2005. "Redistribution and Education Subsidies Are Siamese Twins." *Journal of Public Economics* 89 (11–12): 2005–35.

Chakradhar, Shraddha. 2013. "Striking a Balance on Taxes." *MIT News*, May 22. https://news.mit.edu/2013/student-profile-stefanie-stantcheva-0522.

Cutler, David. 2021. "Interview with Bennett Prize Winner Stefanie Stantcheva." *CSWEP News* 2021 (1). https://www.aeaweb.org/content/file?id=13968.

Edwards, Bruce, and Rhoda Metcalfe. 2025. *Women in Economics*. "Stefanie Stantcheva on Thoughts that Matter." IMF Podcasts, July 15. https://www.imf.org/en/news/podcasts/all-podcasts/2025/07/15/stefanie-stantcheva-women-in-econ.

Farhi, Emmanuel, and Iván Werning. 2013. "Insurance and Taxation over the Life Cycle." *Review of Economic Studies* 80 (2): 596–635.

Henríquez, Marjorie. 2022. "Stefanie Stantcheva: Getting into People's Heads." *Finance and Development Magazine,* September. https://www.imf.org/en/publications/fandd/issues/2022/09/pie-getting-into-people-heads-stefanie-stantcheva.

Jacobs, Bas, and A. Lans Bovenberg. 2011. "Optimal Taxation of Human Capital and the Earnings Function." *Journal of Public Economic Theory* 13 (6): 957–71.

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): 1892–924.

Levitt, Steven D. 2025. *People I (Mostly) Admire*. Episode 165. Produced by Morgan Levey. "The Economist Who (Gasp!) Asks People What They Think." Freakonomics Radio Network, August 29. https://freakonomics.com/podcast/the-economist-who-gasp-asks-people-what-they-think/

Luttmer, Erzo F. P. 2001. "Group Loyalty and the Taste for Redistribution." *Journal of Political Economy* 109 (3): 500–528.

Mirrlees, James A. 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies* 38 (2): 175–208.

Vickrey, William. 1945. "Measuring Marginal Utility by Reactions to Risk." *Econometrica* 13 (4): 319–33.

Journal of Economic Perspectives—Volume 40, Number 3—Summer 2026—Pages 259–266

Recommendations for Further Reading

Timothy Taylor

This section will list readings that may be especially useful to teachers of undergraduate economics, as well as other articles that are of broader cultural interest. In general, with occasional exceptions, the articles chosen will be expository or integrative and not focus on original research. If you write or read an appropriate article, please send a copy of the article (and possibly a few sentences describing it) to Timothy Taylor, preferably by e-mail at <taylort@macalester.edu>, or c/o Journal of Economic Perspectives, Macalester College, 1600 Grand Ave., Saint Paul, MN 55105.

Smorgasbord

Ana Margarida Fernandes and Tristan Reed provide a menu of 15 industrial policy tools, and how and when they are more or less likely to work, in *Industrial Policy for Development: Approaches in the 21st Century* (World Bank, April 2026, https://openknowledge.worldbank.org/entities/publication/9f8098d5-fa1f-4c1b-97b5-f04262818bb3). "Industrial policy—the range of policy tools that governments use to shape what an economy produces rather than leave it to the discretion of markets alone—is back with a vengeance. Despite recent headlines, advanced economies are not the heaviest users of industrial policy. As this report documents, developing economies use it more intensively. New data show that among upper-middle-income economies—those with per capita incomes ranging from US$5,000 to US$14,000—total business subsidies now average 4.2 percent of gross domestic product (GDP),

■ *Timothy Taylor is Managing Editor, Journal of Economic Perspectives, based at Macalester College, Saint Paul, Minnesota.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.20261510.

the highest on record. A review of the most recent national development plans of 183 countries reveals that all countries target growth of at least one industry, and that, on average, low-income countries target 13—more than twice the number in high-income countries. Interest in industrial policy has seldom been higher . . . Deciding which business activities are strategic is perhaps the most difficult and contested topic in industrial policy. As Nobel laureate Paul Krugman remarked about industrial policy in 1983: 'While there is a valid case for targeting grounded in economic theory, the theoretical basis is too complex and ambiguous to be useful given the current state of knowledge'. Of course, the last four decades have seen significant progress in economic measurement, and recent years have seen a resurgence of interest in industrial policy among economists. Nonetheless, Krugman's argument still largely holds."

Christina D. Romer and David H. Romer provide "An early retrospective on monetary policy in the Powell era" (Brookings Institution, Hutchins Center Working Paper #106, June 2026, https://www.brookings.edu/articles/an-early-retrospective-on-monetary-policy-in-the-powell-era/). "To most economists—ourselves included—Chair Powell is a hero. . . . The continuing soundness of the U.S. economy, stability of our financial markets, and respect for the Federal Reserve are due in no small part to his effective leadership. For that we must all be grateful. But that gratitude does not mean that scholars should not evaluate policy in the Powell era with the same rigor and dispassion as they would the tenures of other Fed chairs. . . . [W]e focus on what we see as six relatively distinct policy episodes. These are: (1) the interest rate and balance sheet normalization in 2018 and early 2019; (2) the reversal of both these policies in mid- and late-2019; (3) the aggressive expansionary response to the COVID 19 pandemic in 2020 and early 2021; (4) continued loose policy in 2021 as inflation surged; (5) the rapid tightening in 2022 and 2023 to fight inflation; and (6) the interest rate cuts starting in mid-2024 and severe threats to Fed independence."

In the *Economic Survey 2025–26*, the latest edition of the annual report from India's Ministry of Finance (January 2026, https://www.indiabudget.gov.in/economicsurvey), I was struck by some comments from V. Anantha Nageswaran in the "Preface": "India's export performance since the start of the millennium tells its own story. In general, services exports have outpaced goods exports. In particular, over the five years since 2020, the compounded annual growth rate of total exports has been 9.4%, while that of merchandise exports has been only 6.4%. Services have done much of the heavy lifting, creditable and macro-stabilising, but not a substitute for the goods-based export ecosystems that ultimately underpin durable external and currency stability. The Information Technology-Enabled Services Sector has been India's mainstay for growth and exports since the dawn of the millennium. International experience indicates that while service exports are economically valuable, they do not systematically compel broad upgrades in state capacity, as successful firms can bypass weak institutions, relocate easily, and generate limited economy-wide pressure on governments to reform. Unlike manufacturing exports, they do not impose hard fiscal, employment, or logistical constraints on the State, allowing institutional weakness

to persist even alongside globally competitive firms. So, manufacturing matters." The report later notes: "[S]ervices are increasingly integrated into manufacturing through activities such as design, R&D, logistics, software development, and professional services, reflecting the growing 'servicification' of production systems. This is evident in products such as smart devices, whose value is driven by software ecosystems; medical equipment/wearables bundled with diagnostic and remote-monitoring services; and automobiles, which are increasingly described as 'software on wheels.' As manufacturing becomes increasingly technology and data-intensive, services such as ICT, finance, compliance, and after-sales support account for a growing share of value creation. International experience suggests that this integration is a crucial channel for enhancing value addition, export competitiveness, and employment."

David M. Cutler and Lev Klarnet address "Has the United States bent the health care cost curve?" (*Brookings Papers on Economic Activity*, Spring 2026, https://www.brookings.edu/articles/has-the-united-states-bent-the-health-care-cost-curve). "Aggregating across a number of data sets and analyses, we highlight the role of five central factors in the [health care] spending slowdown. First, technological innovation has become more likely to save money over time. This shows up in medications that prevent acute events and in surgeries that can be performed cheaper and with fewer complications. We estimate the development of cost saving technology accounts for about 21% of the spending slowdown. Second, demand for some types of care has fallen. Demand changes might be due to changes in reimbursement, higher cost sharing, and tighter insurer restrictions on utilization. Together, demand changes accounts for 10–26% of the spending slowdown, with the range reflecting economic uncertainty about effects. Third, long-run supply elasticities tend to be greater than short-run supply elasticities, leading to price reductions over time. Examples of this include pharmaceuticals going off patent and imaging prices declining. These account for 6% of the spending slowdown. Fourth, health status has improved in other ways that we do not understand, but that may be due to reduced smoking and other preventive care. A healthier population needs less care than does a less healthy one. These trends account for 7% of the spending slowdown. Fifth, there was a reduction in the rate of price growth, from 1–2% above general inflation to about general inflation. This results in about 24% of the spending slowdown. . . . Considering our primary question, we conclude that the US has bent the health care cost curve. The role of technology in particular is fundamentally different from what it was in the past, and that means that cost growth has slowed relative to the past. That said, the cost curve has not bent as much as it could, or as much as it needs to."

Yueran Ma, Mengdi Zhang, and Kaspar Zimmermann compile evidence on "Business Concentration Around the World: 1900–2020" (University of Chicago Becker-Friedman Institute for Economics, February 27, 2026, https://bfi.uchicago.edu/insights/business-concentration-around-the-world-1900-2020). "In this paper, we document two sets of facts about the evolution of the organization of production over the past century. These facts hold broadly, across a variety of market-based economies where we can find comprehensive long-run data on the firm size distribution. First, sales, net income, and equity capital have become increasingly

concentrated in the largest firms. In many countries, the largest 1% firms by sales now account for around 80% of economy-wide sales, up from around 50% in the early 20th century. The long-run increases of concentration also hold at the industry level. Second, employment concentration has been relatively stable. The largest 1% firms by employees account for roughly 50% of economy wide employment throughout the 20th century. One exception is retail/wholesale trade, where employment concentration has risen almost as much as sales concentration. These pervasive patterns . . . show that the rising dominance of large firms is a widespread phenomenon, not limited to the recent decades or the United States. Moreover, large firms scale not so much with labor, and possibly more with capital (except in industries like retail where expanding automation has been more challenging thus far)."

Jesse Tack, Jisang Yu, and Roderick M. Rejesus discuss "Recent approaches in agricultural production economics: Where the heck are the prices?" (*Food Policy*, May 2026, https://www.sciencedirect.com/science/article/pii/S0306919226000205). "The purpose of this review is to offer a practitioner's perspective on the evolution of empirical methods since the mid-twentieth century within the context of agricultural production economics. If there was a singular contribution it would rest on our presentation and discussion of research papers that have leveraged more recently developed/popularized empirical approaches outside of the traditional duality-based frameworks. However, we also (i) provide a backward-looking historical context of where these approaches fit within the general methodological landscape dating back to (at least) the 1960s; and (ii) discuss very recent innovations in structural approaches that suggest a path forward in which strengths of both 'old' and 'new' methods are blended together. The intended audience for this contribution is widespread: (a) people outside of the agricultural production economics arena can learn about the historical evolution of methodological approaches as well as important research contributions for a select group of topics; (b) younger researchers within the agricultural production landscape can gain an appreciation for how their approaches fit into the historical context and also begin thinking about how we might further evolve as a profession; and (c) more established researchers can gain an appreciation of how widespread these newer approaches have become and the types of questions they can answer."

Stephan Haggard, Kyoochul Kim, and Munseob Lee discuss techniques of what they call "forensic economics" in "Studying economic black holes: Lessons from North Korea" (*World Development*, May 2026, 201: 107315, https://www.sciencedirect.com/science/article/pii/S0305750X26000045?via%3Dihub). From the abstract: "Some economies are 'black holes' where reliable data is scarce due to government control, low capacity, or conflict. Despite these challenges, researchers have found ways to gather useful information. This paper draws on the literature on North Korea to review six key methods: satellite imagery, reports from aid agencies, trade data, prices, refugee surveys, and official documents. These sources are imperfect, and require close attention to research design and measurement error. Nonetheless, they demonstrate that it is possible to extract information from economic black holes and to draw meaningful insights about them."

Globalization and Its Imbalances

Finance & Development has published a five-paper symposium on "Geoeconomics" (June 2026, https://www.imf.org/en/publications/fandd/issues). Christopher Clayton, Matteo Maggiori, and Jesse Schreger contribute "Understanding Geoeconomics in a Volatile World." "The academic study of geoeconomics dates most prominently to 1945, when economist Albert Hirschman published *National Power and the Structure of Foreign Trade*. In it, he examines how Nazi Germany had structured its economy to maximize leverage over its neighbors during the interwar period. He rejected the naive view that because trade is voluntary and mutually beneficial, it is geopolitically harmless. Benefits can be mutual, Hirschman argues, without being symmetrical. And asymmetry is how power builds. Since Hirschman's time, economists have left the study of global power dynamics largely to political scientists and historians, who have led the development of this area of research. Though almost every economics student encounters the Herfindahl-Hirschman Index, few know it was invented to measure the economic power of nations, not firms. . . . Our work shows that there is a trade-off between gains from trade and economic security. The same mechanisms that are the classic foundations of the gains from trade—economies of scale and specialization—also generate economic dependence. The domestic alternatives that countries did not build up are poor substitutes for globally dominant inputs, such as Chinese manufacturing or US financial services and technology. This lack of alternatives leaves the countries exposed to coercion. As the global economy increasingly relies on goods and services that have strategic complementarities and economies of scale, these mechanisms are likely to increase in importance."

Steven A. Altman and Caroline R. Bastian survey the extent of globalization in the *DHL Global Connectedness Report 2026* (https://www.dhl.com/global-en/microsites/core/global-connectedness/report.html). "The DHL Global Connectedness Index does not indicate a shift from international to domestic activity across trade, capital, information, and people flows. Global connectedness reached a record high in 2022 and has not changed appreciably through 2025. . . . U.S. tariff increases only modestly reduced forecast global trade growth. Other countries supported trade growth by not raising tariffs, and many negotiated new trade deals to secure access to alternative markets. . . . The world remains far from a split into disconnected geopolitical blocs. Only 4–6% of global goods trade, greenfield FDI, and cross-border M&A have shifted away from geopolitical rivals over the past decade. Trade flows shifted more toward neutral countries than to close allies, implying more 'de-risking' than 'friendshoring'. . . . Prominent narratives about deglobalization are driven more by politics and public policy than by actual shifts in cross-border flows. While the risk of deglobalization has risen and the pattern of connectedness is shifting, the world overall remains as connected as ever."

The IMF offers a guide to "Understanding Global Imbalances" (April 5, 2026, https://www.imf.org/en/publications/policy-papers/issues/2026/04/06/understanding-global-imbalances-575234). "Global imbalances have been a recurrent feature of the global economy since the 1870s and countries have

switched positions over time . . . The United Kingdom, for example, ran sustained surpluses prior to World War II—reflecting colonial trade and large investment income—before shifting into a structural deficit position after 1945. More recently, the US transitioned from surplus to deficit in the 1970s, while Germany and Japan moved into surplus. . . . Four economies—the US, China, Germany, and Japan—account for roughly two-thirds of global imbalances. The US deficit—equivalent to 4 percent of GDP as of 2024—has been financed by capital inflows and portfolio investors seeking dollar assets. Germany and Japan's surpluses have been driven by high saving rates, joined by substantial surpluses in China beginning in the 2000s. Oil-exporting countries' surpluses fluctuate with commodity prices, creating episodic contributions to global imbalances. . . . While current account surpluses and deficits can be appropriate when they reflect economic fundamentals and desirable policies, the buildup and persistence of large imbalances raise concerns when they are driven by policy distortions and unwind in a disorderly manner."

Hélène Rey, Beatrice Weder di Mauro, and Jeromin Zettelmeyer have edited a collection of 17 essays in *Paris Report 4: The New Global Imbalances* (Centre for Economic Policy Research, 2026, https://cepr.org/publications/books-and-reports/p398). From the introductory essay by di Mauro and Zettelmeyer: "In sum, global current account imbalances reflect domestic saving–investment gaps. They can support growth when financed sustainably and directed toward productive uses, but they become risky when large, persistent, and tied to rising leverage or asset bubbles. What matters for these risks is not bilateral trade balances but the underlying macroeconomic conditions. Durable adjustment therefore requires domestic policy changes, not trade measures alone. . . . The ideal adjustment would involve the main systemic economies—at least the United States, China, and Europe—rebalancing simultaneously and in a coordinated manner. Such an approach would reduce the risk that adjustment in one economy simply shifts imbalances elsewhere or triggers destabilising spillovers. In simple terms, the required policy mix is well known. The United States would raise national saving, primarily through credible fiscal consolidation, thereby reducing its reliance on external financing. China would lower excess saving by rebalancing toward household consumption—strengthening social safety nets, boosting disposable income, and shifting away from investment- and export-led growth. Europe, for its part, would increase investment, particularly in infrastructure, defence, and the green transition, thereby absorbing more domestic and global savings."

Interviews

Erika McEntarfer was Commissioner of the US Bureau of Labor Statistics until August 1, 2025. Neale Mahoney discusses the experience with her and what comes next for federal statistics in "The hidden backbone: The data behind the economy" (Stanford Institute for Economic Policy Research, "Econ to Go," March 26, 2026, https://siepr.stanford.edu/av/Econ-To-Go-podcast-hidden-backbone-data-us-economy).

On her interaction with the Department of Government Efficiency, McEntarfer says: "I actually had a whole basket of AI related projects when DOGE arrived early in the administration. The early word was they were gonna help us with AI. And I was like, 'Great, we could use some more resources here.' So, you know, I had this whole list of projects for them and instead I wound up sitting across the table from a member of DOGE and they were like, 'So we want to fire all of these statisticians and replace them with the AI.' And I was like, 'I don't think that's actually possible, but if you can explain to me how it is possible, I am all ears.' And then they would just stare at me blankly and tell me that I was not cooperating with their vision. I was like, 'No, I don't actually understand how you replace a time series statistician with an AI model, but if you can explain it to me, I'm all ears.'"

Tim Sablik interviews "Ellen McGrattan: On measuring what businesses do, developing effective tax policy, and searching for answers beyond the lamppost" (*Econ Focus*: Federal Reserve Bank of Richmond, First/Second Quarter 2026, https://www.richmondfed.org/publications/research/econ_focus/2026/q1-q2_interview). On total factor productivity (TFP), McGrattan says: "In some sense, I've been struggling with trying to look inside the black box of TFP all my career. . . . My interest in business cycles partly stems from trying to measure what goes into TFP. . . . What is TFP? It's what we don't know, it's the part that we need to fill in. It's not just some magic dust that's in the air. . . . If you buy, say, a computer, it has to be put on your balance sheet. But if you're a dentist and you spend time building your patient list, that's not put on any balance sheet. That patient list is the thing you sell when you retire or relocate, and that asset contributes to the value in the business, but we never see it until it gets sold or transferred somehow. There are 40 million active businesses in the United States, and most have assets that we can't see. Assets like customer bases or trademarks—until there's a transaction, we can't see them. You might have a good accounting system that you developed within your business, or you're a chef and you have recipes, and we can only see those things if you trade them. But most ongoing businesses don't list these assets on a balance sheet, so we never get to see them. And that's why we need a theory to infer it. . . . It all ties back to work I was doing to measure movements in the economy over the business cycle, but now I would say the bigger issue is how to measure all activity in the U.S. economy."

Discussion Starters

Virginia Postrel tells the story of "Engineering the disposable diaper: Benjamin Spock told mothers in the mid-twentieth century to buy six dozen cloth diapers and a covered pail. Within a decade, both were obsolete" (*Works in Progress*, April 24, 2026, https://worksinprogress.co/issue/engineering-the-disposable-diaper/). "After buying Charmin Paper Company in 1957, Procter & Gamble began looking for ideas for new paper products. Motivated by the less pleasant aspects of spending time with his new grandchild, the company's director

of exploratory development, Victor Mills, suggested disposable diapers. After analyzing existing products and conducting consumer research, P&G created a dedicated diaper research group. The research this group conducted, like that of its successors and competitors, wasn't glamorous. It didn't advance basic science. It wasn't even an obvious route to profit. . . . It was a high-stakes gamble that required solving difficult engineering problems. How that happened represents the kind of hidden progress that leads to everyday abundance."

Anton Korinek and Patrick McKelvey ask "Where is AI in GDP Statistics (Peterson Institute for International Economics, May 2026, https://www.piie.com/publications/policy-briefs/2026/where-ai-gdp-statistics). "We estimate . . . that nominal AI compute spending grew by more than 140 percent per year each in 2024 and 2025, raw compute capacity by more than 200 percent per year, and quality-adjusted AI output by more than 2,000 percent per year. The divergence between this picture of the AI economy and the one drawn by conventional GDP statistics is itself an informative macroeconomic signal. Treating the AI sector as a coherent economic entity in its own right yields a preliminary estimate of nominal AI GDP of roughly $250 billion in 2025—comparable in size to the US scheduled passenger airline industry—yet growing at approximately 2,600 percent per year in quality-adjusted terms. We argue that US statistical agencies and economic policymakers should start now to assemble better data on AI activity in AI satellite accounts—focused subsets of the national economic statistical accounts . . . They should begin to incorporate AI productive-capacity measures into medium-term projections and scenario analysis."

The American Economic Association

Correspondence relating to advertising, business matters, permission to quote, or change of address should be sent to the AEA business office: aeainfo@vanderbilt.edu. Street address: American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. For membership, subscriptions, or complimentary access to JEP articles, go to the AEA website: http://www.aeaweb.org. Change of address notice must be received at least six weeks prior to the publication month.

The Journal of
Economic Perspectives

Summer 2026, Volume 40, Number 3

Symposia

Artificial Intelligence

Tariffs

Instrumental Variables

Article

Feature

AMERICAN ECONOMIC ASSOCIATION